THE PROCEEDINGS OF THE PHYSICAL SOCIETY

FROM JANUARY 1947 TO NOVEMBER 1947

VOLUME 59

MDCCCLXXIV

Published by
THE PHYSICAL SOCIETY
1 Lowther Gardens, Prince Consort Road,
London S.W. 7

Printed by
TAYLOR AND FRANCIS, LTD.,
Red Lion Court, Fleet Street, London E.C. 4

OFFICERS AND COUNCIL, 1947-48

PRESIDENT

G. I. Finch, M.B.E., D.Sc., F.R.S.

VICE-PRESIDENTS
who have filled the Office of President

C. H. Lees, D.Sc., F.R.S.
Sir Frank Smith, G.C.B., G.B.E., D.Sc., F.R.S.
Sir Owen Richardson, M.A., D.Sc., F.R.S.
W. H. Eccles, D.Sc., F.R.S.
A. O. Rankine, O.B.E., D.Sc., F.R.S.
The Right Hon. Lord Rayleigh, M.A., Sc.D., F.R.S.
T. Smith, M.A., F.R.S.
Allan Ferguson, M.A., D.Sc.
Sir Charles Darwin, K.B.E., M.C., M.A., Sc.D., F.R.S.
E. N. da C. Andrade, Ph.D., D.Sc., F.R.S.
D. Brunt, M.A., Sc.D., F.R.S.

VICE-PRESIDENTS

Sir Edward Appleton, G.B.E., K.C.B., D.Sc., F.R.S.
A. J. Philpot, C.B.E., M.A., B.Sc.
H. R. Robinson, D.Sc., Ph.D., F.R.S.
W.D. Wright, D.Sc.

HONORARY SECRETARIES

W. Jevons, D.Sc., Ph.D. (*Business*)
H. H. Hopkins, Ph.D. (*Papers*)

HONORARY FOREIGN SECRETARY

E. N. da C. Andrade, Ph.D., D.Sc., F.R.S.

HONORARY TREASURER

H. Shaw, D.Sc.

HONORARY LIBRARIAN

R. W. B. Pearse, D.Sc., Ph.D.

ORDINARY MEMBERS OF COUNCIL

W. S. Stiles, D.Sc., Ph.D.
C. H. Collie, M.A., B.Sc.
J. D. Bernal, M.A., F.R.S.
D. Roaf, M.A., D.Phil.
A. C. G. Menzies, M.A., D.Sc.
R. Peierls, C.B.E., M.A., Dr.Phil., D.Sc., F.R.S.
F. C. Toy, C.B.E., D.Sc.
J. H. Awbery, M.A., B.Sc.
L. F. Bates, D.Sc., Ph.D.
R. C. Evans, M.A., Ph.D.
L. C. Martin, D.Sc.
C. E. Wynn-Williams, D.Sc., Ph.D.

COLOUR GROUP

Chairman
J. G. Holmes, B.Sc.
Honorary Secretary
W. D. Wright, D.Sc.

OPTICAL GROUP

Chairman
L. C. Martin, D.Sc.
Honorary Secretary
E. W. H. Selwyn, B.Sc.

LOW-TEMPERATURE GROUP

Chairman
Sir Alfred Egerton, M.A., Sec.R.S.
Honorary Secretary
G. G. Haselden, Ph.D.

ACOUSTICS GROUP

Chairman
H. L. Kirke, C.B.E., M.I.E.E.
Honorary Secretaries
W. A. Allen, B.Arch., A.R.I.B.A.
A. T. Pickles, O.B.E., M.A.

SECRETARY-EDITOR

(*from* 1 *September* 1947)
Miss A. C. Stickland, Ph.D.
1 Lowther Gardens, Prince Consort Road, London S.W.7
(*Telephone*: KENsington 0048)

CONTENTS

Part 1. 1 *January* 1947

PAGE
Portrait of Professor D. BRUNT, M.A., Sc.D., F.R.S., President of the Physical Society, 1945–47 . . . . *frontispiece*
A. C. MERRINGTON and E. G. RICHARDSON. The break-up of liquid jets . . 1
DAVID OWEN. The lines of force through neutral points in a magnetic field . . 14
P. VIGOUREUX. Sensitivity and impedance of electro-acoustic transducers . . 19
H. FRÖHLICH and R. A. SACK. Light absorption and selective photo-effect in adsorbed layers 30
Corrigendum 33
W. B. LEWIS, F.R.S. Fluctuations in streams of thermal radiation . . . 34
C. R. BURCH, F.R.S. Reflecting microscopes 41
C. R. BURCH, F.R.S. Semi-aplanat reflecting microscopes 47
S. PATERSON. The heating or cooling of a solid sphere in a well-stirred fluid . . 50
Sir EDWARD APPLETON, F.R.S. and W. J. G. BEYNON. The application of ionospheric data to radio communication problems : Part II 58
E. F. DALY and G. B. B. M. SUTHERLAND. An infra-red spectroscope with cathode-ray presentation 77
J. C. JAEGER. Equivalent path and absorption in an ionospheric region . . 87
W. J. G. BEYNON. Oblique radio transmission in the ionosphere, and the Lorentz polarization term 97
H. G. HOWELL. On the spectra of CS and CSe 107
F. A. B. WARD. A simple optical model demonstrating the principle of the Bragg x-ray spectrometer 111
F. A. B. WARD. A mechanical model illustrating the uranium chain reaction . 113
DEREK J. PRICE. The emissivity of hot metals in the infra-red 118
DEREK J. PRICE. The temperature variation of the emissivity of metals in the near infra-red 131
A. R. UBBELOHDE. The freezing-in of nuclear equilibrium 139
M. L. OLIPHANT, F.R.S. Rutherford and the modern world 144
DISCUSSION on paper by E. W. H. SELWYN and J. L. TEARLE entitled " The performance of aircraft camera lenses (*Proc. Phys. Soc.*, **58**, 493 (1946)) 155
Reviews of books 157

Part 2. 1 *March* 1947

R. WEIL. A note on the intermittency effect 161
C. GURNEY. Delayed fracture in glass 169
B. BLEANEY, J. H. N. LOUBSER and R. P. PENROSE. Cavity resonators for measurements with centimetre electromagnetic waves 185
H. N. V. TEMPERLEY. The behaviour of water under hydrostatic tension : III . 199
H. A. ELLIOTT. An analysis of the conditions for rupture due to Griffith cracks . 208
J. BROSSEL. Multiple-beam localized fringes : Part I.—Intensity distribution and localization 224
J. BROSSEL. Multiple-beam localized fringes : Part II.—Conditions of observation and formation of ghosts 234

PAGE
J. C. Evans. The determination of thermal lagging times 242
F. R. N. Nabarro. Dislocations in a simple cubic lattice 256
Maurice Milbourn. A note on the effect at the cathode of an arc between copper electrodes 273
R. F. Hanstock. Damping capacity, strain hardening and fatigue 275
J. M. Cowley and A. L. G. Rees. Refraction effects in electron diffraction . . 287
O. Klemperer. Electron optics and space charge in strip-cathode emission systems 302
T. Smith, F.R.S. A series for the stationary value of a function 323
Reviews of books 326

Part 3. 1 *May* 1947

L. F. Bates and A. S. Edmondson. The adiabatic temperature changes accompanying the magnetization of cobalt in low and moderate fields 329
E. Glückauf. Investigations on absorption hygrometers at low temperatures . 344
G. G. Macfarlane. A theory of flicker noise in valves and impurity semiconductors 366
D. K. C. MacDonald and R. Fürth. Spontaneous fluctuations of electricity in thermionic valves under retarding field conditions 375
R. Fürth and D. K. C. MacDonald. Statistical analysis of spontaneous electrical fluctuations 388
Discussion on the foregoing papers by G. G. Macfarlane (p. 366) and R. Fürth and D. K. C. MacDonald (pp. 375 and 388) 403
F. C. Frank. The mass of the neutrino 408
R. E. Peierls. What experiments are needed in fundamental physics ? . . . 412
B. Bleaney and R. P. Penrose. Collision broadening of the inversion spectrum of ammonia at centimetre wave-lengths. I.—Self-broadening at high pressure . 418
R. C. Brown. The fundamental concepts concerning surface tension and capillarity 429
E. E. Vago and R. F. Barrow. Ultra-violet absorption band-systems of PbO, PbS, PbSe and PbTe 449
D. V. Gogate and P. D. Pathak. The Landau velocity of liquid helium II . . 457
Sir Edward Appleton, F.R.S. and R. Naismith. The radio detection of meteor trails and allied phenomena 461
Mary P. Lord, A. L. G. Rees and M. E. Wise. The short-period time variation of the luminescence of a zinc sulphide phosphor under ultra-violet excitation . 473
Addendum to Discussion on the paper by R. F. Schmid and L. Gerö entitled "Photo-chemical decomposition of CO" (*Proc. Phys. Soc.* **58,** 701 (1946)) . 502
Obituary notices :
  Sir James Jeans, O.M., F.R.S. 503
  Thomas Howell Laby, M.A., Sc.D., F.R.S. 506
  George Blackford Bryan, O.B.E., D.Sc., M.I.E.E. 508
  William Barron Coutts, M.A., B.Sc. 508
Reviews of books 509

Part 4. 1 *July* 1947

W. J. G. Beynon. Some observations of the maximum frequency of radio communication over distances of 1000 km. and 2500 km. 521

PAGE

DISCUSSION on papers by Sir E. APPLETON and W. J. G. BEYNON, " The application of ionospheric data to radio communication problems " (*Proc. Phys. Soc.*, **59**, 58 (1947)); W. J. G. BEYNON, " Oblique radio transmission in the ionosphere and the Lorentz polarization term " (*Phys. Proc. Soc.*, **59**, 97 (1947)); and W. J. G. BEYNON, this part, p. 521 534

A. E. BATE and M. E. PILLOW. Mean free path of sound in an auditorium . . 535

T. B. RYMER and C. C. BUTLER. Determination of the crystal structure of gold leaf by electron diffraction 541

R. DONALDSON. A colorimeter with six matching stimuli 554

N. E. G. HILL. The recognition of coloured light signals which are near the limit of visibility 560

N. E. G. HILL. The measurement of the chromatic and achromatic thresholds of coloured point sources against a white background 574

E. A. NEUMANN. A time micrometer of high accuracy 585

J. G. HOLMES. Colorimetry in the glass industry 592

S. P. SINHA. Ultra-violet bands of Na_2 610

W. D. ALLEN and J. L. SYMONDS. Experiments in multiple-gap linear acceleration of electrons 622

C. GURNEY. Thermodynamic relations for two phases containing two components in equilibrium under generalized stress 629

E. G. DYMOND. The Kew radio sonde 645

M. L. OLIPHANT, J. S. GOODEN and G. S. HIDE. The acceleration of charged particles to very high energies. 666

J. S. GOODEN, H. H. JENSEN and J. L. SYMONDS. Theory of the proton synchrotron 677

G. WYLLIE. The hole theory of diffusion 694

W. B. MANN and L. G. GRIMMETT. The Imperial College high-voltage generator . 699

E. WOLF and W. S. PREDDY. On the determination of aspheric profiles . . 704

Corrigenda 711

Reviews of books 711

Part 5. 1 *September* 1947

D. BRUNT, F.R.S. Some physical aspects of the heat balance of the human body . 713

MAX JAKOB. Some investigations in the field of heat transfer 726

J. D. CRAGGS and W. HOPWOOD. Ion concentrations in spark channels in hydrogen. 755

J. D. CRAGGS and W. HOPWOOD. Electron/ion recombination in hydrogen spark discharges 771

ROBERT WEIL. The variation of the reflectivity of nickel with temperature . . 781

H. O. W. RICHARDSON. Magnetic focusing between inclined plane pole-faces . 791

H. CRAIG. The production of a uniform magnetic field over a specific volume by means of twin conducting circular coils 804

H. G. W. HARDING and R. B. SISSON. Distribution coefficients for the calculation of colours on the C.I.E. trichromatic system for total radiators at 1500–250–3500° K., and 2360° K. ($C=14\,350$) 814

J. W. DUNGEY and CATHERINE R. HULL. Coaxial electron lenses 828

T. SMITH, F.R.S. Note on aplanatic lenses for unit magnification 844

J. S. HEY, S. J. PARSONS and F. JACKSON. Reflexion of centimetric electro-magnetic waves over ground, and diffraction effects with wire-netting screens . . 847

J. S. HEY and G. S. STEWART. Radar observations of meteors 858

PAGE

C. M. G. LATTES, P. H. FOWLER and P. CUER. A study of the nuclear transmutations of light elements by the photographic method 883

DISCUSSION on paper by E. F. DALY and G. B. B. M. SUTHERLAND entitled " An infra-red spectroscope with cathode-ray presentation " (*Proc. Phys. Soc.*, **59**, 77 (1947)) 901

Corrigenda 901

Reviews of books 902

Part 6. 1 *November* 1946

R. E. SIDAY. The optical properties of axially symmetric magnetic prisms : Part 1. 905

J. HAMILTON. The theory of radiation damping 917

W. J. BATES. A wavefront-shearing interferometer 940

R. C. FAUST and S. TOLANSKY. A transparent-replica technique for interferometry. 951

C. DOMB. The theory of an oscillator coupled to a long feeder, with applications to experimental results for the magnetron 958

F. HOYLE. On the formation of heavy elements in stars 972

R. LATHAM. Nuclear magnetic moments 979

A. S. EDMONDSON. The matching of high-frequency transmission lines using a frequency-variation method 982

J. P. ANDREWS. Thermoelectric power of cadmium oxide 990

A. L. G. REES. The calculation of potential-energy curves from band-spectroscopic data 998

A. L. G. REES. Note on the interpretation of the visible absorption spectrum of bromine 1008

D. F. RUSHMAN and M. A. STRIVENS. The effective permittivity of two-phase systems 1011

D. G. AVERY and R. WITTY. Diffusion pumps : a critical discussion of existing theories 1016

R. EISENSCHITZ. The effect of temperature on the thermal conductivity and viscosity of liquids 1030

Discussions 1036

Obituary notices :

- FRIEDERICH PASCHEN 1040
- PAUL LANGEVIN 1041
- JOHN HENRY STRONG 1042

Reviews of books 1042

Index to Volume 59 1049

Index to Reviews of books, Volume 59 1056

Proceedings at the meetings, Session 1946–47 viii

Report of Council for the year 1946 xvii

Report of the Honorary Treasurer for the year 1946 xxi

PROCEEDINGS AT THE MEETINGS OF THE PHYSICAL SOCIETY

SESSION 1946-47

15 *May* 1946*

The fourth meeting of THE LOW-TEMPERATURE GROUP, in the Department of Chemistry Technology, Imperial College, London S.W. 7. Sir Alfred Egerton was in the Chair.

The High-pressure and Low-temperature Laboratories were visited, and the following lectures were delivered:

" Storage of liquefied gases ", by Sir Alfred Egerton and M. Pearse;

" Liquefaction in relation to gas manufacture ", by T. A. Hall.

* The record of this meeting was inadvertently omitted from the *Proceedings*, **58**, page ix (November 1916).

13 *September* 1946

The twenty-fourth meeting of THE OPTICAL GROUP, at Imperial College, London S.W. 7. Instr.-Capt. T. Y. Baker was in the Chair.

The following papers were read and discussed:

" The performance of aircraft camera lenses ", by E. W. H. Selwyn and J. L. Tearle;

" The effect of the angle of incidence of the exposing light rays upon the resolving powers of photographic materials ", by J. M. Gregory;

" The Ross 25-inch *F*/6·3 Express E.M.I. lens ", by O. G. Hay and G. A. Richmond.

3 *October* 1946

The twenty-ninth meeting of THE COLOUR GROUP, at the Lighting Service Bureau, Electric Lamp Manufacturers' Association, London W.C. 2. Dr. R. K. Schofield was in the Chair.

The following papers were read and discussed:

" The recognition of coloured light signals which are near the limit of visibility ", by N. E. G. Hill;

" The measurement of the chromatic and achromatic thresholds of coloured point sources against a white background ", by N. E. G. Hill.

3 *October* 1946

Science Meeting, at the Science Museum, London S.W. 7. The President, Professor D. Brunt, was in the Chair.

The following were elected to Fellowship, the last four being transferred from Student Membership: Raymond Ernest Adlington, Arthur Jerrard Booth, William John Rawson Calvert, Gordon James Chamberlin, Rowland Braddock Clayton, Alan Hugh Cook, Pierre Fleury, Mohamed Gharid Abd El-Galil, René Léon Gauthier, Brian Michael Green, Kariamanikkam Srinivasa Krishnan, Polidoor August Cyriel Mortier, Josiah Percy Reed, Charles Sadron, Arthur Henry Sully; John Geoffrey Dawes, Arthur Dennis Kent, Bernard John Perrett, Fáthi Asad Qaddura.

It was announced that the Council had elected the following to Student Membership: John Frederick Archard, Bernard George Childs, Geoffrey Howard Cockett, Alfred William Crook, Leo Jung, Richard Peter Martin, Alfred Maurice Mendoza, Colin John Moore, Walter Eric Spear, Stanley Alfred Tibbs, Richard Wilson, Dennis Ralph Workman.

The thirtieth Guthrie Lecture was delivered by Professor Max Jakob (Chicago), who took as his subject " Some investigations in the field of heat transfer ".

7 *October* 1946

Science Meeting, at the Royal Institution, London W. 1. The President, Professor D. Brunt, was in the Chair.

The following were elected to Fellowship, the last three being transferred from Student Membership: Robert James Benzie, Jacob Clay, Walter Henry Cole, Allen Dinsdale, Joseph William Fisher, Mahmoud Hessaby, James Stanley Hey, David Barrable Hodges, Arthur Magnus John Janser, Enaf Morrice Job, Otto Klemperer, Wilfred Bennett Lewis, Aubrey William Pryce, George Ormerod Rawstron, Darcy Walker, Charles George Webb, Jerzy Wyhowski; Kenneth Robert Atkins, Robert Frank Crocombe, Wilfrid John Mellors.

The third Rutherford Memorial Lecture was delivered by Professor M. L. E. Oliphant (Birmingham), whose subject was " Rutherford and the modern world ".

13 *November* 1946

The fifth meeting of THE LOW-TEMPERATURE GROUP, at the Science Museum, London S.W. 7. Sir Alfred Egerton was in the Chair.

A discussion on " The cultivation of a thermodynamic outlook " was held, the opening paper being by Sir Charles Darwin.

The Meeting was preceded by the first *Annual General Meeting of the Low-Temperature Group* for the presentation of the Committee's report for 1946 and for the election of Officers and Committee for 1946–47.

15 *November* 1946

Science Meeting, at the Science Museum, London S.W. 7. Mr. T. Smith, Past-President, was in the Chair.

Arthur John King and Walter William Wright were elected to Fellowship.

It was announced that the Council had elected the following to Student Membership: Maurice John Bott, David St. Pierre Bunbury, John Michael Tate Clark, David Ian Clough, Alan Cyril Coates, Brian Anthony Collett, Julian Cedric Cook, Colin Armstrong Crofts, David Elwyn Davies, Rudolph Dehn, Kenneth Henry Dixey, James Percival Dixon, James Edmund Gibbs, Richard Henry Eldridge, Geoffrey Charles Fletcher, John William Charles Gates, Harold William Greensmith, Alec Geoffrey Hester, Peter Edward Hodgson, Ronald Michael Horsley, Peter Andrew Reay Kinghorn, Atley Mortimer Lawrence, Harry Malbon, Leonard Mandel, Thomas Alan Margerison, Peter Frederick Mariner, Maurice George Milford, George Desmond Morgan, John Harold Denis Morris, Michael Louis Nunn, Arthur Beck Parker, Evan Thomas de la Perrelle, Hubert Lloyd Roberts, Dennis William Sciama, Richard Ingram Shewell, Josef Kazimierz Skwirzynski, John Smith, Peter David Southgate, Priscilla Halford Strange, Alan Howard Ward, George Alan Wilkins, Marjorie Elizabeth Woolcock, Kenneth Herbert Reginald Wright.

A demonstration of neon stroboscopic lamps was given by L. F. Berry (Ferranti Ltd.).

The following papers were read and discussed:

" Electron optics and space charge in strip-cathode emission systems ", by O. Klemperer;

" The application of ionospheric data to radio communication problems ", by Sir Edward Appleton and W. J. G. Beynon;

" Oblique radio transmission in the ionosphere and the Lorentz polarization term ", by W. J. G. Beynon;

" Some observations of the maximum frequency of radio communication over distances of 1000 km. and 2500 km.", by W. J. G. Beynon.

22 *November* 1946

The twenty-fifth meeting of THE OPTICAL GROUP, at Imperial College, London S.W. 7. Instr.-Capt. T. Y. Baker was in the Chair.

A paper on " The diffraction theory of the phase-contrast test " was read by E. H. Linfoot.

Demonstrations of new phase-contrast microscopes were given by E. W. Taylor and A. Hughes, the latter of whom also showed two phase-contrast films.

6 *December* 1946

Science Meeting, at the Science Museum, London S.W. 7. The President, Professor D. Brunt, was in the Chair.

The following were elected to Fellowship: Robert Wirenfeldt Asmusson, Edwin Richard Collins, Leonard George Grimmett, Edward Hubert Linfoot, James William McHugo (transferred from Student Membership), Francis Hugh Campbell Morgan Minnis, Barry Joseph Cornelius O'Sullivan, Philip Litherland Teed, John Cecil Thomas, George Leonard Turney.

It was announced that the Council had elected the following to Student Membership: Norman Adams, Douglas Russell Bennett, Albert John Bird, Richard Arthur Brown, Robert Gordon Cawthorne, Ronald William Cornish, Laurie Cotton, Michael Daniels, Philip Davis, John Sydney Dugdale, David Noel Ferguson Dunbar, Herbert Sigmund Eisner, Robert Donald Bruce Fraser, Peter Gay, Geoffrey Shakleton Hawkes, John E. Hooper, Neil Howells, Donald Gordon Neal Hunter, Robert Edward Jones, Mary Eveline Manson, Norman Owen Matthews, R. Furley Mellor, Leo Melzer, Peter William Mummery, Terence Edwin Olver, Kenneth Robert Pallant, Alfred John Parker,David Ambrose Edward Roberts, John Francis Lloyd Roberts, William Kenneth Roberts, Stanley George Saint, Max Smoulevitz, Geoffrey Frank Snelling, Ganapathy Srikantia Shrikantia, Eric Swindlehurst, Desmond Philip Cameron Thackeray, John Ronald Thomas, Peter Ewart Watson, James Norwood Whyte, G. N. Wood, Raymond William Henry Wright.

A lecture on " Fundamental concepts concerning surface tension and capillarity " was delivered by R. C. Brown, and was followed by an informal discussion.

19 *December* 1946

The thirtieth meeting of THE COLOUR GROUP, in the Library of the Royal Society of Arts, Adelphi, London W.C. 2. Dr. R. K. Schofield was in the Chair.

The recently published Report on *Defective Colour Vision in Industry* was introduced by W. D. Wright and H. M. Cartwright and afterwards discussed.

20 *December* 1946

Science Meeting, at the University of Birmingham. The President, Professor D. Brunt, was in the Chair.

The following were elected to Fellowship: Cedric John Brown, Claude James Bradish (transferred from Student Membership), William Harold Joseph Childs, Karl George Emeléus, Derek Henry Fender, Hamilton Hartridge, Leo Kowarski.

A Conference on " Fundamental problems in modern physics " was held, the following papers being read and discussed:

" What experiments are needed in fundamental physics ? ", by R. E. Peierls;

" The design of proton synchrotrons ", by J. S. Gooden;

" Some practical aspects of synchrotron design ", by D. W. Fry;

" Recent experiments on highly ionized gases ", by J. Sayers (read by P. B. Moon).

" Extra-terrestrial nuclear reactions ", by F. Hoyle.

The cyclotron and synchrotron under construction in the Nuffield Laboratory were inspected.

10 *January* 1947

Science Meeting, at the Science Museum, London S.W. 7. The President, Professor D. Brunt, was in the Chair.

The following were elected to Fellowship: William Douglas Allen, Ernst Billig, Percy Hatfield, George Pearce (transferred from Student Membership), Arkadiusz Piekara, John Eric Roberts.

It was announced that the Council had elected the following to Student Membership: Bernarr Francis Atherton, Adeline Mary Dale, Edwin Roland Dobbs, Peter Derek Fochs, Norman Frank Godwin, Lewis John Griffiths, Graham Hulse, Frederick Arthur Jacobs, Geoffrey Knight, David Ross Knowles, Sheila Bernice Lataste, Edward Thornton Linacre, John Howard McGuire, Gordon Eric Mawer, Eric Rees Pearcey, Lloyd Julian Perper, Patrick Reginald Dan Pomeroy, Michael Shepherdson, Alec Geoffrey Thompson, Kenneth Alan Turner, Frank Lewis Ward, Eric Roy Wooding.

The twenty-third (1946) Duddell Medal was presented to Karl Weissenberg, who afterwards gave a lecture on his work on x-ray goniometers.

The second (1946) Charles Vernon Boys Prize was presented to R. W. Sutton, who afterwards gave a brief account of his work on receiving valves and cathode-ray tubes (local oscillators and the skiatron) for radio and radar.

22 *January* 1947

The sixth meeting of THE LOW-TEMPERATURE GROUP, at the Science Museum, London S.W. 7. Sir Alfred Egerton was in the Chair.

A lecture on " The separation of oil gases " was delivered by M. Ruhemann.

29 *January* 1947

The thirty-first meeting of THE COLOUR GROUP, at the Lighting Service Bureau of the Electric Lamp Manufacturers' Association, London W.C. 2. Dr. R. K. Schofield was in the Chair.

A lecture on " The use of colour in factories " was delivered by H. D. Murray and was followed by an informal discussion.

31 *January* 1947

Science Meeting, at the Science Museum, London S.W. 7. The President, Professor D. Brunt, was in the Chair.

The following were elected to Fellowship: Cecil Benjamin Allsopp, Harold Frederick Cook, Richard Heinrich Herz, Christopher Guy Ardern Hill (transferred from Student Membership), Ernst Jacobus Marais, Desmond Ramsay Rexworthy, John Howard Richards.

It was announced that the Council had elected the following to Student Membership: Donald Geoffrey Avery, John Sidney Blakemore, Roger Alec Bones, Peter Edward Douglas, Robert Lockhart Graham, Dennis John Hine, Geoffrey Munday, Gordon Arthur John Orchard, Walter Steckelmacher, Peter Millner Stott.

The following papers were read and discussed:

" The radio detection of meteor trails and allied phenomena ", by Sir Edward Appleton and R. Naismith;

" Radar observations of meteors ", by J. S. Hey and G. S. Stewart;

" A study of transient radar echoes from the ionosphere, including observations made during the solar eclipse of July 1945 ", by E. Eastwood and K. A. Mercer;

" Radio echoes from meteors ", by A. C. B. Lovell, C. J. Banwell and J. A. Clegg.

A gramophone record of " whistling meteors " was heard, with a description by G. R. M. Garrett.

14 *February* 1947

The twenty-sixth meeting of THE OPTICAL GROUP, at Imperial College, London S.W. 7. Professor L. C. Martin was in the Chair.

Reports on the Réunion d'Opticiens held in Paris in October 1946 were presented and discussed:

" Physical optics and technical applications ", by L. V. Chilton, H. H. Hopkins and J. S. Preston;

" Visual photometry, colorimetry and physiological optics ", by W. S. Stiles.

19 *February* 1947

Inaugural meeting of THE ACOUSTICS GROUP, in the Jarvis Hall of the Royal Institute of British Architects, London W. 1. Mr. H. L. Kirke was in the Chair.

A draft constitution of the Group was discussed and adopted. The Chairman, Vice-Chairman, Secretaries and Committee of the Group for 1947–48 were elected.

An address on " The contribution of acoustical science to allied studies " was delivered by Dr. Alex Wood.

21 *February* 1947

Science Meeting, at the Science Museum, London S.W. 7. The President, Professor D. Brunt, was in the Chair.

The following were elected to Fellowship, the last fifteen being transferred from Student Membership: Thomas Edward Allibone, Montague Ernest Clarkson, Georges Louis Charles Léon Dejardin, Horace Edward Dohoo, Patrick Joseph Doyle, Charles Geoffrey Blythe Garrett, Anna Modayil Mani, Thomas Harold Oddie, Charles Robert Oswin, James Eldred Skewes, Norman Thompson, Meredith Wooldridge Thring, Reginald Harbert Wadie; Alan Barker, Bruce Alexander Bilby, Michael Davis, Eric Foster, Robert Louis Gordon, Ralph Percy Hudson, John Fletcher Hutton, Raymond Douglas Lowde, Gordon Kingsley Monks, Geoffrey Harwood Moss, Ralph William Nicholls, Kenneth John Veryard, Bernard Miles Wheatley, Garth Angus Wheatley, Peter Stephen Williams.

The following papers were read and discussed:

" Spontaneous fluctuations of electricity in thermionic valves under retarding field conditions ", by D. K. C. MacDonald and R. Fürth;

" Statistical analysis of spontaneous electrical fluctuations ", by R. Fürth and D. K. C. MacDonald;

" Theory of flicker noise in valves and impurity semi-conductors ", by G. G. Macfarlane.

26 *February* 1947

The seventh meeting of THE LOW-TEMPERATURE GROUP, at the Science Museum London S.W. 7. Sir Charles Darwin was in the Chair.

A lecture on " The graphical and geometrical representation of thermodynamic functions " was delivered by N. R. Kuloor.

19 *March* 1947

Science Meeting, at the University of Manchester. Professor P. M. S. Blackett was in the Chair.

A Conference was held on " Meteors, comets and the radio detection of meteors", the following papers were read and discussed:

" Meteors ", by P. J. M. Prentice;

" Comets ", by J. G. Porter;

" Experimental work on the radio echoes from meteors ", by A. C. B. Lovell;

" The theory of meteor ionization ", by N. Herlofson.

The Conference was followed on 20 March by a visit to the Jodrell Bank Experimental Station of the University.

19 *March* 1947

The eighth meeting of THE LOW-TEMPERATURE GROUP, at the Clarendon Laboratory, Oxford. Sir Charles Darwin was in the Chair.

A survey of the work of the Low-Temperature Department of the Laboratory was given by Professor F. Simon, and was followed by a tour of the Laboratory and short discussions of selected topics.

21 *March* 1947

The twenty-seventh meeting of THE OPTICAL GROUP, at Imperial College, London S.W. 7. Professor L. C. Martin was in the Chair.

The following papers were read and discussed:

" The physical properties of glass ", by W. C. Hynd;

" The degree of inhomogeneity of optical glass and some aspects of its bearing on optical performance ", by T. H. Wang.

26 *March* 1947

The thirty-second meeting of THE COLOUR GROUP, in the Lecture Room of The Royal Photographic Society, London S.W. 7. Mr. J. G. Holmes was in the Chair.

A paper on " A statistical investigation of some aspects of colour harmony ", by M. E. Clarkson, O. L. Davies and T. Vickerstaff was read and discussed.

The meeting was preceded by the sixth *Annual General Meeting of The Colour Group*, for the presentation of the Committee's report on the work of the Group in 1946–47 and for the election of Officers and Committee for 1947–48.

28 *March* 1947

Extraordinary General Meeting, at Imperial Collge, London S.W. 7. The President, Professor D. Brunt, was in the Chair.

The following Special Resolution was passed unanimously:

> That the Articles of Association be altered by the substitution of the following Article for the present Article 36:
>
> > " 36. Any Fellow may at any time compound for all annual subscriptions thereafter to become due from him. The composition fee shall be the product of the amount of the annual subscription payable by such Fellow and the factor appropriate to the age of such Fellow at his last birthday before his giving notice of desire to compound, such factor being determined according to the following Table:—

TABLE

Age	Factor	Age	Factor	Age	Factor	Age	Factor
21	30	31	25	41	19·5	51	12·5
22	29·5	32	24·5	42	19	52	12
23	29	33	24	43	18	53	11
24	28·5	34	23·5	44	17·5	54	10
25	28	35	23	45	17	55	9·5
26	27·5	36	22·5	46	16	56	8·5
27	27	37	22	47	15·5	57	7·5
28	26·5	38	21·5	48	15	58	7
29	26	39	21	49	14	59	6
30	25·5	40	20	50	13·5	60	5 "

28 *March* 1947

Science Meeting, at Imperial College, London S.W. 7. The President, Professor D. Brunt, was in the Chair.

The following were elected to Fellowship, the last seven being transferred from Student Membership: Constance E. Arregger, John Oscar Guy Barrett, Robert Benedict Bourdillon, Donald William Fry, Reinhold Fürth, John Lloyd Jones, William Noel Sproson, Alexander Stewart; Kenneth Henry Clarke, George William Greenlees, Oliver Samuel Heavens, Ronald Mayoh, Kenneth Elmslie Munn, Derek John Price, Richard Wallace Whorlow.

It was announced that the Council had elected the following to Student Membership: Sultana Z. Ali, Frank Arthur Chappell, Charles John Gravett, John Stephen Halliday, John Douglas Jolley, George Harry King, Har Nath, James Robert Rogers, David Wilkie.

The following papers were read and were followed by a short discussion on the fracture of solids:

"Delayed fracture in glass", by C. Gurney;

"Extension of Griffith's theory of rupture to three dimensions", by R. A. Sack.

23 *April* 1947

Science Meeting, at Imperial College, London S.W. 7. Professor E. N. da C. Andrade, Foreign Secretary and Past President, was in the Chair.

The following were elected to Fellowship, the last fourteen being transferred from Student Membership: Harendra Kumar Acharya, Benjamin Chapman Browne, Archibald Hugh Campbell, John Drummond Craggs, Colin Allen Haywood, Lachlan Mackinnon, Kurt Sitte, Frank Alan Underwood, Harish Chandra Verma; Peter Andrews, George Cowper, James Maurice Daniels, Brian Philip Day, Stephen Gerhard Friedrich Frank, John Hodgkinson, Alfred Ernest Kay, Thomas Rundell Lomer, James Nicol, Charles Arthur Padgham, George Sparling Parry, Eric Robinson, Mary Lois Tiffany (*née* Joyce), Edith Eva Vago.

It was announced that the Council had elected the following to Student Membership: Raymond Alfred Allen, Edward James Burge, James Charles Ezekiel Button, Werner Freitag, Michael James Hart, Frank Elsden Neale.

A discussion on "The spark discharge" was held, the opening paper being read by L. B. Loeb (Berkeley, California), and other papers by T. E. Allibone, J. M. Meek and J. D. Craggs.

8 *May* 1947

Annual General Meeting, at Imperial College, London S.W. 7. The President, Professor D. Brunt, was in the Chair.

The minutes of the two General Meetings, held on 16 May 1946, were read and confirmed.

The Reports of the Council and the Honorary Treasurer and the Annual Accounts for 1946 were adopted.

The Officers and Council and the Auditors for 1947–48 were elected.

Votes of thanks were accorded to the Rector and Governing Body of Imperial College and Sir George Thomson, the Managers of the Royal Institution, the Director of the Science Museum, and The Electric Lamp Manufacturers' Association for excellent accommodation at meetings; to the Royal Commission for the Exhibition of 1851 and Dr. Evelyn Shaw for the office and library accommodation at 1 Lowther Gardens, Prince Consort Road, London S.W. 7; to Dr. A. F. C. Pollard for preparing the U.D.C. Index Slips for the *Proceedings*; and to the retiring Officers and Council.

8 *May* 1947

Science Meeting, at Imperial College, London S.W. 7. The newly elected President, Professor G. I. Finch, was in the Chair.

The following were elected to Fellowship: John Flavell Coales, William Evans Hugh Humphreys, Harold Lister Kirke, James Stewart McPetrie, Egon Orowan, Albert Shaw, John Harry Walrond Simmons, Robert Allan Smith, Robert Tucker, Victor Frederick George Tull, Walter Weinstein.

The retiring President, Professor D. Brunt, delivered his address on "Some aspects of the heat balance of the human body".

9 *May* 1947

The twentiy-eighth meeting of THE OPTICAL GROUP, at Imperial College, London S.W. 7. Professor L. C. Martin was in the Chair.

A paper on "Photographic resolving power of lenses" was read by Mr. E. W. H. Selwyn, and was followed by an informal discussion.

The meeting was preceded by the fifth *Annual General Meeting of The Optical Group*, for the presentation of the Committee's report on the work of the Group in 1946–47 and for the election of the Officers and Committee for 1947–48.

30 *May* 1947

Science Meeting, at Imperial College, London S.W. 7. The President, Professor G. I. Finch, was in the Chair.

The following were elected to Fellowship, the first two being transferred from Student Membership: Robert Alan Croft, John Gorham; Willis Jackson, Frederick Wilson Jones, Donald Burgess McNeill, Ronald James Post, Edward Eric Shelton, Jack Vennart.

It was announced that the Council had elected the following to Student Membership: Alan George Clegg, David Walter Evans, Joseph Abraham Franks, John Lewis Fyson, Walter Grattidge, John Hampson, Kenneth Gerald Hinton, Michael Harold Kreps, John Rowland Mallard, William Alan Runciman, Norman Chester Underwood.

The following papers were read and discussed:

"Thermoelectric power of cadmium oxide", by J. P. Andrews;

"The optical properties of axially symmetric magnetic prisms", by R. E. Siday;

"Coaxial electron lenses", by J. W. Dungey and Miss C. R. Hull.

4 *June* 1947

The thirty-third meeting of THE COLOUR GROUP, at the Lighting Service Bureau of the Electric Lamp Manufacturers' Association, London W.C. 2. Mr. J. G. Holmes was in the Chair.

A paper on "The colour temperature of light sources" was read by H. G. W. Harding, and was followed by an informal discussion.

26 *June* 1947

The first meeting of THE ACOUSTICS GROUP, at the Royal Institute of British Architects, London W. 1. Mr. H. L. Kirke was in the Chair.

A symposium on "Sound absorption and reverberation" was held, the opening papers of the morning and afternoon sessions being:

"Panel absorbents of the Helmholtz type", by P. V. Bruel (Gothenberg, Sweden);

"Reverberation time as an index of room performance", by J. Moir.

27 *June* 1947

Extraordinary General Meeting, at Imperial College, London S.W. 7. The President, Professor G. I. Finch, was in the Chair.

The President made a statement on the Council's proposals to raise the annual subscriptions in 1948 and to replace "The Proceedings of the Physical Society" by "The Magazine of Physics", and submitted the following Special Resolution:

That the Articles of Association of the Society be altered as follows:

(i) Article 11, by substituting
  (*a*) in line 1, for the word "two" the word "three";
  (*b*) in line 4, for the words "one guinea a year" the words "two guineas a year in accordance with Article 31";
  (*c*) in line 6, for the words "the Proceedings" the words "The Magazine of Physics".
(ii) Article 31, by substituting
  (*a*) in line 2, for the words "two guineas or one guinea" the words "three guineas or two guineas"; and

(*b*) in lines 11 and 12, for the words " two guineas or may pay an annual subscription of one guinea " the words " three guineas or may pay an annual subscription of two guineas ".

(iii) Article 42, by substituting for the words " ten shillings and sixpence " the words " fifteen shillings ".

(iv) Article 44, by substituting for the words " the Proceedings " in line 1 the words " The Magazine of Physics ".

(v) Article 81, by substituting for the words " the Proceedings of the Society " in line 4 the words " The Magazine of Physics ".

After a discussion, in which an amendment to raise the Fellowship subscription to four guineas instead of three guineas was defeated, clauses (i) (*a*) (*b*), (ii) and (iii) were passed unanimously, and clauses (i) (*c*), (iv) and (v) were withdrawn by the Chairman.

---

27 *June* 1947

Science Meeting, at Imperial College, London S.W. 7. The President, Professor G. I. Finch, was in the Chair.

The following were elected to Fellowship: James Wynne Dungey (transferred from Student Membership), Eric Harold Harden, David William Hills, Cyril Alfred Hogarth, Harold George Jerrard, Harry Jones, William George Kennings Kilbourn, Stanley Lucas, Norman Veall.

It was announced that the Council had elected Gerd Martin Nathan and Deryck Arthur John Walliker to Student Membership.

A demonstration of a model representing the action of a full-wave rectifier was given by D. J. Morgan.

A lecture on " Applications of multiple-beam interferometry to surfaces and thin films " was delivered by S. Tolansky, and was followed by an informal discussion.

REPORT OF COUNCIL FOR THE YEAR ENDED 31 DECEMBER 1946

INTRODUCTORY AND GENERAL

The year under review was marked by the resumption of the Society's Exhibition of Scientific Instruments and Apparatus, by the first presentation of the Holweck Prize, by a highly successful International Conference, and by an unprecedented increase of the Fellowship. Unfortunately, however, there was a large financial deficit—a sharp reminder that the present rates of annual subscription are far too low in comparison with the services rendered by the Society.

Comparison of this Report with that for 1945 will reveal a marked increase in the volume of the Society's publications ; the barrier which prevents further expansion is lack of paper, caused not only by the inadequacy of the official allowance but also by the difficulty of obtaining it from the manufacturers.

The Council acknowledges with gratitude the continued generosity of the Royal Commission for the Exhibition of 1851 and its kindly Secretary, Sir Evelyn Shaw, in providing the accommodation at 1 Lowther Gardens. The re-equipment and re-furnishing of the rooms have been continued, though by no means completed.

The Council also records its thanks to the Managers of the Royal Institution, to the Director of the Science Museum in whose lecture theatres most of the Science Meetings were held, and to the Rector and Governing Body of Imperial College and Professors Sir George Thomson and H. V. A. Briscoe for the great privilege of holding the Exhibition in the Physics and Chemistry Departments of the College.

MEETINGS

An Annual General Meeting and an Extraordinary General Meeting were held at the Royal Institution on 16 May 1946, the former for the presentation of the Reports of the Council and the Honorary Treasurer for 1945 and the election of the Officers and Council for 1946–47, and the latter for the temporary suspension of Article 36.

Eleven Science Meetings, in addition to those of the three Groups, were held during the year, six at the Science Museum, three at the Royal Institution, one at Imperial College and one in the Physics Department of the University of Birmingham.

The Birmingham Meeting, the first of a new series of meetings to be held at Universities outside London, was the occasion of a successful Conference on " Fundamental Problems in Modern Physics ". One of the London meetings was held jointly with the Royal Meteorological Society for a Conference on " Meteorological Factors in Radio-Wave Propagation ", a report of which will be published during the Summer of 1947. Of the other nine meetings in London, three were devoted to Papers and Demonstrations, two to Lecture Surveys in continuation of the series initiated during the War and four to Special Lectures and functions noted in the following paragraphs (see Guthrie Lecture, Rutherford Lecture and Holweck Prize).

EXHIBITION OF SCIENTIFIC INSTRUMENTS AND APPARATUS

The Society's 30th Exhibition at Imperial College was held on 1–3 January 1946 and was opened by the Rt. Hon. Sir Stafford Cripps, President of the Board of Trade. At that time, within a few months of the end of the war, the conditions for the preparation and organization of the Exhibition could hardly have been less favourable to the exhibitors, to the Society's small office staff, or to the printers.

Nevertheless, the high reputation established by the pre-war Exhibitions was maintained or even surpassed. The Exhibition is undoubtedly a function of acknowledged pre-eminence in this country, and its Catalogue is widely acclaimed as a valuable work of reference. The attendance was unexpectedly and embarrassingly large—nearly double that in 1938 or 1939. A first edition of the Catalogue was disposed of within a few hours of the opening, and, in view of the heavy demand for copies both in this country and overseas, a second edition was produced later in the year. Entirely new problems of organization have been presented to the Exhibition Committee, and the work for the 31st (1947) Exhibition was started in good time.

INTERNATIONAL CONFERENCE

By arrangement with Sir Lawrence Bragg, an International Conference on " Fundamental Particles and Low-Temperature Physics " was held at the Cavendish Laboratory, Cambridge, on 22–27 July 1946. The attendance was so large that many members of the Society who wished to attend were unable to obtain accommodation and admittance. Those present included guests and visitors from Canada, U.S.A., U.S.S.R., India, China and most European countries. The opening address was by Professor Niels Bohr (Copenhagen). One day was devoted to papers on, and an inspection of, the work of the Cavendish and Royal Society Mond Laboratories and to the ceremonial opening of the Austin Wing by the Rt. Hon. Sir John Anderson. The papers read and discussed at the Conference are to be published as two Special Reports of the Society, which should be ready towards the end of 1947.

MEMORIAL LECTURES AND AWARDS

Guthrie Lecture

At the Science Museum on 3 October Professor Max Jakob (Illinois Institute of Technology, Chicago) delivered the 30th Guthrie Lecture, the subject of which was " Some investigations in the field of heat transfer " (*Proceedings*, **59** (1947), p. 726).

Rutherford Memorial Lecture

The 2nd (1944) Rutherford Lecture, postponed on account of the long absence of Professor J. D. Cockcroft from this country during and after the War, was delivered by him at the Science Museum on 8 February 1946, the title being " Rutherford : life and work after the year 1919 ; with personal reminiscences of the Cambridge period " (*Proceedings*, **58** (1946), p. 625).

The 3rd (1946) Lecture was delivered at the Royal Institution on 7 October by Professor M. L. E. Oliphant, who took as his subject " Rutherford and the modern world " (*Proceedings*, **59** (1947), p. 144).

Duddell Medal and Charles Vernon Boys Prize

The 23rd (1946) Duddell Medal was awarded to Dr. Karl Weissenberg (Shirley Institute of the Cotton Research Association) in recognition of his work on x-ray goniometers, and the 2nd (1946) Charles Vernon Boys Prize to Mr. R. W. Sutton (Services Electronics Research Laboratory, Admiralty) for his work on receiving valves and cathode-ray tubes (local oscillators and the skiatron) for radio and radar. The presentations were unavoidably postponed to a date in 1947.

Holweck Prize

The Council made the first (1946) award of the Prize to Professor Charles Sadron (University of Strasbourg) in recognition of his work on the mechanical and optical properties of liquids, and had the pleasure of welcoming to London for the presentation ceremony Professor P. Fleury and Professor G. A. Boutry, as representatives of the Société Française de Physique, and Mme Sadron. The valuable assistance of the British Council in the accommodation and entertainment of the Society's guests is gratefully acknowledged. In addition to founding the Médaille Holweck, to be awarded to each prize-winner, the Société Française has presented to the Physical Society a bronze plaque of Fernand Holweck for the Council Room. The Officers and Council record their great appreciation of this generous gesture by their French colleagues.

PUBLICATIONS

Proceedings

During the year the official allowance of paper for the *Proceedings* was increased to the amount used annually before the war. The larger quantity enabled the Society to increase the size of Vol. **58** (1946) and the number of copies printed, but it is quite inadequate to the needs of a Society which has doubled its membership and its outside subscribership in less than a decade. It has been decided to discontinue the publication of Universal Decimal Classification Index Slips at the end of Vol. **58** (1946), as the information given in the Slips is now readily available in *Physics Abstracts*. The Council again thanks Professor A. F. C. Pollard for his valuable services in the preparation of the Slips since they were first supplied in 1937.

In connexion with the International Conference reported above, a meeting of physicists from European countries brought out clearly the difficulty felt on the Continent owing to the lack of a journal in which short papers on new work could receive rapid publication. Discussions have been in hand during the year to consider whether the nature, and indeed the title, of the Society's *Proceedings* could be modified so as to meet this demand. A definite decision will be reached and announced during 1947.

REPORTS ON PROGRESS IN PHYSICS

The long-awaited Volume **10** (1944–45) and Volume **4** (1937, reprinted) appeared in September 1946, but the number of copies delivered by the end of the year was far short of the large number of orders already entered and still being received. Volumes **7**, **8** and **9** were in very heavy demand and the stock of each was almost entirely disposed of by the end of the year. The office of the American Institute of Physics has continued to help the Society by supplying copies of the *Reports* to American physicists. The preparation of Volume **11** (1946–47) was begun under the General Editorship of Dr. W. B. Mann and was transferred to a new Editorial Board upon his departure to Chalk River, Ontario.

SPECIAL REPORTS

The *Report on Defective Colour Vision in Industry*, on the preparation of which a Sub-Committee of the Colour Group was engaged for several years, appeared towards the end of 1946, and is in great demand. The congratulations and thanks of the Society are due to the members of the Committee upon the successful completion of its valuable work. Mention has already been made of other Special Reports which are in active preparation, namely, those on the Conferences on Meteorological Factors on Radio-Wave Propagation, Fundamental Particles and Low-Temperature Physics.

REPRESENTATION ON OTHER BODIES

The following appointments of representatives in 1946 are reported :—

National Committee for Physics : Profressor G. I. Finch, Professor M. L. E. Oliphant, Mr. R. S. Whipple.

National Committee for Scientific Radio : Mr. R. Naismith, Mr. C. W. Oatley.

Committee of Management of Science Abstracts : Mr. J. H. Awbery, Dr. A. G. Gaydon, Dr. A. C. Menzies, Dr. D. Roaf.

Committee for the Jubilee of the Discovery of the Electron : Dr. J. H. Brinkworth, Dr. W. Jevons, Dr. H. Shaw.

Celebration (*in Paris*, 29–31 *October* 1946) *of the Jubilee of the Discovery of Radioactivity* : Professor H. R. Robinson.

OBITUARY

The Council records with regret the deaths of the following Fellows :—Mr. J. L. Baird, Professor G. B. Bryan, Mr. J. B. Butler Burke, Mr. W. B. Coutts, Captain E. A. Hoghton, Sir James Jeans, Professor T. H. Laby, Professor P. Langevin, Dr. J. J. Manley, Mr. R. W. D. Mayall, Professor H. C. Plummer (Honorary Fellow of the Optical Society), Dr. M. Poser, Dr. J. Schofield and Mr. J. J. Steward; and the death of a Student Member, Mr. R. E. Hickman. The Society was represented by the President at the funeral of Sir James Jeans, and by Professor J. B. Bernal at the ceremony of homage to Professor Langevin in Paris.

MEMBERSHIP

Roll of Membership		Hon. Fellows	Hon. Fellows, Optical Society	*Ex-officio* Fellows	Fellows	Student Members	Total
Totals, 31 Dec. 1944		10	1*	4	1167	275	1457
Changes during 1946	Newly elected				139	53	
	Transferred				46 } 185	46	
	Deceased		1		13	1 } 57	
	Resigned				6	8	
	Lapsed				1 } 20	2	
	Net increase		−1		165	−4	160
Totals, 31 Dec. 1946		10	0	4	1332	271	1617

As the above table shows, the sixteen-hundred mark was passed during the year. The increase in the Fellowship and in the total membership was more than double that in 1945 and far greater than in any previous year ; it is satisfactory to note that the increase was largely accounted for by new elections to Fellowship, and the transfer from Student Membership was not far behind the record for 1945, while the new elections to Student Membership just failed to make up for the large transfer. The relevant facts for the last seven years are as follows :—

Year	1940	1941	1942	1943	1944	1945	1946
Net increase in total Membership	−14	6	107	73	63	75	160
Net increase in Fellowship	−25	37	45	34	55	80	165
Newly elected Fellows	20	28	57	42	56	56	139
Transfer from Student Membership	6	13	20	22	23	52	46

GROUPS

COLOUR GROUP

At the Sixth Annual General Meeting of the Group, which was held at the Science Museum on 13 March 1946, Dr. R. K. Schofield and Dr. W. D. Wright were re-elected as Chairman and Honorary Secretary, respectively, and the Committee for 1946–47 was elected.

Five Science Meetings took place in 1946, and are briefly recorded in the *Proceedings.**

The *Report on Defective Colour Vision in Industry,* prepared by one of the two Sub-Committees, was published in December 1946 and was the subject of a Discussion at one of the Science Meetings and of a broadcast talk by Dr. Wright. It has received excellent press notices and the sales of copies in this country and overseas are highly satisfactory.

The report of the other Sub-Committee, on Colour Terminology, is nearly ready for the printers.

OPTICAL GROUP

The Group held its Fifth Annual General Meeting at Imperial College on 26 April 1946, when Instr.-Capt. T. Y. Baker and Mr. E. W. H. Selwyn were re-elected as Chairman and Honorary Secretary and the Committee for 1946–47 was elected.

During the year six Science Meetings were held, particulars of which are given in the *Proceedings.** Several of the papers read at the meetings have been published in the *Proceedings* † and others will appear in due course.

LOW-TEMPERATURE GROUP

During 1946, the first year of the Group's activities, four Science Meetings were held, as noted in the *Proceedings.**

The Second Annual General Meeting of the Group took place at the Science Museum on 13 November 1946, when the Officers and Committee for 1946–47 were appointed ; Sir Alfred Egerton and Sir Charles Darwin were re-elected as Chairman and Vice-Chairman, respectively, and Mr. G. G. Haselden was elected as Honorary Secretary.

MEMBERSHIP

The membership of the three Groups on 31 December 1946 was as follows :—

	Colour	Optical	Low-Temperature
Members of the Physical Society	101	191	38
Members of participating bodies	74	73	33
Members of subscribing firms	17	105	
Other Members	16	12	14
Totals	208	381	85

---

* *Proceedings,* **58** (1946), pp. viii–xi; and **59** (1947), pp. viii–x.

† *Proceedings,* **58** (1946), pp. 65, 493, 759, 769; and **59** (1947), pp. 41, 560, 574.

REPORT OF THE HONORARY TREASURER FOR THE YEAR ENDED 31 DECEMBER 1946

For the first time since 1938 the accounts show an unwelcome excess of expenditure over income. The total expenditure (£10,424) and the total income (£8940) were higher than in 1945 by £5473 and £3620 respectively.

The large increase (£3695) in the cost of publications is due to several causes: the expansion of the *Proceedings* (in the size of the volume and in the number of copies printed), the production of two volumes of *Reports on Progress in Physics* (the reprinted Vol. IV and the new Vol. X) and the *Report on Defective Colour Vision*, and the rising costs of printing and paper. Administration expenditure (£2898) was double that in 1945, but this is not surprising in view of the post-war expansion in the activities of the Society and the necessary increase of office staff and equipment. Expenses in connection with meetings (which now include the cost of printing for meetings, hitherto shown as a separate item) show a relatively small increase (£77).

Purchases of office furniture and equipment during the year were met partly by the Herbert Spencer Bequest, which has now been fully spent, and partly from the general fund.

There are considerable increases of income from all four of the main sources (subscriptions, sales, advertisements and dividends), but they have not kept abreast of the rising expenditure.

The large net deficit (£1484) is rather misleading, since the accounts do not include the value of the stock of publications, which was abnormally large at the end of the year. Nevertheless, comparison with the satisfactory balances in the accounts for 1944 and 1945 clearly indicates the need for a larger annual income; this can only be secured by increases of annual subscriptions, retail sale prices and advertisement rates.

A publications grant of £450 by the Royal Society is gratefully acknowledged. Donations amounting to £1840 for the International Conference at Cambridge in July 1946 are also gratefully acknowledged; the balance of £639 shown in the Balance Sheet will defray part of the cost of producing the Reports of the Conference.

Investments increased in total value by £1937, of which £1000 was a new investment in 3% Defence Bonds and the remaining £937 represents appreciation of earlier investments.

(Signed) H. SHAW,
Honorary Treasurer.

21 *March* 1947.

SPECIAL FUNDS

W. F. STANLEY TRUST FUND

	£	s.	d.		£	s.	d.
Carried to Balance Sheet	259	0	0	£300 Southern Railway Preferred Ordinary Stock	199	0	0
				£442 Southern Railway Deferred Ordinary Stock	60	0	0
	£259	0	0		£259	0	0

UDDELL MEMORIAL TRUST FUND

CAPITAL

	£	s.	d.		£	s.	d.
Carried to Balance Sheet	374	0	0	£400 3½% War Loan Inscribed "B" Account	374	0	0

REVENUE

	£	s.	d.		£	s.	d.
Honorarium	20	0	0	Balance on 31 December 1945		15	9
Medal and Certificate	6	5	0	Interest on War Loan	14	0	0
Balance carried to Balance Sheet	1	0	9	Grant from General Fund	12	10	0
	£27	5	9		£27	5	9

SPECIAL FUNDS (*contd.*)

HERBERT SPENCER LEGACY

	£	s.	d.		£	s.	d.
Expenditure on Furniture and Equipment during the year	85	4	3	Balance on 31 December 1945	85	4	3

CHARLES CHREE MEDAL AND PRIZE FUND

CAPITAL

	£	s.	d.		£	s.	d.
Balance carried to Balance Sheet	1865	16	4	Balance on 31 December 1945	1865	16	4

REVENUE

	£	s.	d.		£	s.	d.
Balance carried to Balance Sheet	81	17	0	Balance on 31 December 1945	12	19	10
				Interest on Investments	68	17	2
	81	17	0		£81	17	0

CHARLES VERNON BOYS PRIZE FUND

CAPITAL

	£	s.	d.		£	s.	d.
Balance carried to Balance Sheet	900	0	0	£1132 16*s.* 10*d.* 2½% Consols "B" Account	900	0	0

REVENUE

	£	s.	d.		£	s.	d.
Balance on 31 December 1945	17	7	1	Interest on Investment	28	6	4
Prize	26	5	0	Balance carried to Balance Sheet	16	10	9
Certificate	1	5	0				
	£44	17	1		£44	17	1

HOLWECK PRIZE FUND

	£	s.	d.		£	s.	d.
Prize	100	0	0	Balance on 31 December 1945	887	4	11
Expenses	39	14	0	Interest on Investment	24	1	3
Balance carried to Balance Sheet	796	12	2	Grant from General Fund	25	0	0
	£936	6	2		£936	6	2

ADDENBROOKE BEQUEST

CAPITAL

	£	s.	d.		£	s.	d.
Balance carried to Balance Sheet	337	0	0	£384 6*s.* 7*d.* 2½% Consols "D" Account	337	0	0

REVENUE

	£	s.	d.		£	s.	d.
Book Plates	13	3	6	Balance on 31 December 1945	9	12	0
Balance carried to Balance Sheet	6	0	6	Interest on Investment	9	12	0
	£19	4	0		£19	4	0

LIFE COMPOSITION FUND ON 31 DECEMBER 1946

	£	s.	d.
19 Fellows paid £10	190	0	0
1 Fellow paid £15	15	0	0
17 Fellows paid £21	357	0	0
57 Fellows paid £31 10*s.* 0*d.*	1795	10	0
	£2357	10	0

INCOME AND EXPENDITURE ACCOUNT FOR THE YEAR ENDED 31 DECEMBER 1946

1945 £	EXPENDITURE	£ s. d.	£ s. d.
	To		
801	" Science Abstracts "		798 5 9
2105	" Proceedings "	3901 17 6	
21	" Reports on Progress in Physics "	1755 17 7	
	Special Publications	163 8 4	5821 3 5
335	Postage on Publications and General Correspondence		562 17 1
93	Agenda Papers and Notices		264 13 4
95	Expenses at Meetings		
52	Honoraria to Special Lecturers		78 15 0
	Administration Expenses :		
1373	Secretarial and Clerical Assistance and Office Expenses		2800 19 7
	Grant to Holweck Prize Fund	25 0 0	
75	Grant to Duddell Memorial Fund	12 10 0	37 10 0
	Furniture and Office Equipment	144 14 10	
	Less transfer from Herbert Spencer Legacy	85 4 3	59 10 7
869	*Balance, being excess of Income over Expenditure, carried to General Fund.*		
£5819			£10,423 14 9

1945 £	INCOME	£ s. d.	£ s. d.
	By		
	Subscriptions :		
78	Entrance Fees	128 2 0	
2175	Fellows	2509 13 9	
144	Student Members	184 5 0	
29	Colour Group	26 5 0	
50	Optical Group	70 17 0	
	Low-Temperature Group	11 19 0	
142	For " Science Abstracts " and Advance Proofs	180 7 0	3111 8 9
	Sales :		
1328	" Proceedings "	1663 3 6	
495	" Reports on Progress in Physics "	1148 6 1	
89	Special Publications	339 2 4	3150 11 11
448	Advertisements in " Proceedings "		1061 19 8
371	Dividends from Investments		396 11 5
20	Transfer from Life Composition Fund of amounts paid by Fellows now deceased		84 0 0
450	Publications Grant by the Royal Society		450 0 0
	30th Exhibition (balance)		685 3 8
	Balance, being excess of Expenditure over Income, carried to General Fund		1483 19 4
£5819			£10,423 14 9

BALANCE SHEET AS ON 31 DECEMBER 1946

LIABILITIES

	£	s.	d.	£	s.	d.
Sundry Creditors				1194	13	10
Life Compositions :						
As on 31 December 1945	1906	0	0			
Add Payments during year	535	10	0			
	2441	10	0			
Less Transfer to Income and Expenditure Account.	84	0	0			
				2357	10	0
Subscriptions received in advance :						
Members	54	2	6			
Publications	141	9	8			
				195	12	2
SPECIAL FUNDS :						
W. F. Stanley Trust Fund	259	0	0			
Duddell Memorial Trust Fund	375	0	9			
Charles Chree Medal and Prize Fund	1947	13	4			
Charles Vernon Boys Prize Fund	883	9	3			
Holweck Prize Fund	796	12	2			
Addenbrooke Bequest	343	0	6			
				4604	16	0
GENERAL FUND :						
As on 31 December 1945	7757	10	6			
Less Balance of Income and Expenditure Account	1483	19	4			
				6273	11	2
Cambridge Conference Balance, carried forward				639	2	2
				£15,265	5	4

ASSETS

	£	s.	d.	£	s.	d.
Investments at market value on 31 December 1939 or cost:—						
W. F. STANLEY TRUST FUND :						
£300 Southern Railway Preferred Ordinary Stock.	199	0	0			
£442 Southern Railway Deferred Ordinary Stock.	60	0	0			
				259	0	0
DUDDELL MEMORIAL FUND :						
£400 3½% War Stock				374	0	0
CHARLES CHREE MEDAL AND PRIZE FUND :						
£784 4% Funding Stock	839	0	0			
£1500 2½% Consols	1027	10	0			
				1866	10	0
CHARLES VERNON BOYS PRIZE FUND :						
£1132 16*s.* 10*d.* 2½% Consols				900	0	0
HOLWECK PRIZE FUND :						
£775 3% Defence Bonds				775	0	0
ADDENBROOKE BEQUEST :						
£384 6*s.* 7*d.* 2½% Consols				337	0	0
GENERAL FUND :						
£2100 2½% Consols	1437	10	0			
£1750 3½% War Stock	1636	0	0			
£500 4% Funding Stock	535	0	0			
£400 3% Lancaster Corporation	364	0	0			
£399 L.M.S. 4% Debentures	355	0	0			
£1000 L.M.S. 4% Preference Stock	595	0	0			
£500 L.N.E.R. 4% Debentures	412	0	0			
£150 S.R. 5% Debentures	171	0	0			
£2100 3% Defence Bonds	2100	0	0			
£1000 3% Savings Bonds	1000	0	0			
				8605	10	0
* Market value on 31 December 1946 : £17,701.				13,117	0	0*
Dividends due on Investments				123	17	2
Inland Revenue—Income Tax recoverable for 1946				51	1	4
Subscriptions due				54	5	0
Sundry Debtors				621	0	3
Stock of Paper and Binding Material				227	18	6
31*st Exhibition* (1947) :—*Expenditure to date*				23	17	0
Cash in Post Office Savings Bank Account	984	18	3			
Cash at Bank	14	17	0			
Cash in hand	46	10	10			
				1046	6	1
				£15,265	5	4

We have audited the above Balance Sheet and have obtained all the information and explanations we have required. We have verified the bank balances and the Investments. In our opinion such Balance Sheet is properly drawn up so as to exhibit a true and correct view of the state of the Society's affairs according to the best of our information and the explanations given to us and as shown by the books of the Society.

KNOX, CROPPER & Co.,
Chartered Accountants

SPENCER HOUSE, SOUTH PLACE, E.C. 2.
20th March 1947.

[Photo Elliot and Fry

PROFESSOR D. BRUNT, M.A., Sc.D., F.R.S.,
President of the Physical Society, 1945–47.

[Frontispiece

THE PROCEEDINGS OF THE PHYSICAL SOCIETY

Vol. 59, Part 1 1 January 1947 No. 331

THE BREAK-UP OF LIQUID JETS

By A. C. MERRINGTON and E. G. RICHARDSON,
King's College, Newcastle-upon-Tyne

MS. received 10 *April* 1946

ABSTRACT. The mean drop size produced when a jet of liquid issuing from a nozzle into the atmosphere, breaks up, has been investigated and experiments have been carried out both with fixed nozzles and with moving nozzles (attached to an aircraft) discharging backwards. A relation expressing drop size in terms of viscosity and the relative velocity between the jet and the surrounding air, applicable to both cases, is given.

At low velocities this relation ceases to apply and drop size reaches a limiting value. Factors influencing this limiting figure are discussed and information has also been obtained on the stability of large drops.

Measurements made on the length of the jet confirm the work of Tyler and Richardson but also disclose a hitherto unsuspected anomaly in the case of very viscous liquids.

Similar results were obtained when a jet issued into another liquid with which it did not mix, but showed the effect of the relative viscosity of the liquids on the break-up.

PART I—LIQUID INTO AIR

§ 1. INTRODUCTION

When a jet of liquid issues from a nozzle into the atmosphere, the jet eventually breaks up into drops under the action of disturbances of its equilibrium figure. The jet is observed to break up at a point which can be determined accurately, as long as the pressure behind the nozzle remains constant. While some previous work, both theoretical and experimental, exists on the length of this continuous portion of the jet, little quantitative work has been done on the size and behaviour of the drops subsequent to the break-up. In particular, there is a lack of experimental data on the effect of viscosity on the break-up of jets.

In the main there are three types of break-up. In the first the interplay of inertia and surface tension results in the jet becoming varicose. Rayleigh (1879) examined the conditions under which axial-symmetrical oscillations set up near the nozzle might increase in amplitude. In an inviscid jet he showed that a disturbance having a wave-length 4·4 times the diameter of the jet should grow fastest and eventually break up the jet into drops. Tyler and Richardson (1925) have taken photographs of this type of break-up on capillary jets. Rayleigh subsequently (1892) modified his theory to take account of the viscosity of the liquid, which naturally reduces the rate of growth of the optimum disturbance, whose wave-length in relation to the diameter of the jet remains unchanged.

In the second type of disturbance, the jet becomes sinuous and the resistance of the air to the passage of the humps becomes of more importance than surface tension. This type of break-up has been dealt with mathematically by Weber (1931) but is only amenable to semi-empirical treatment.

Since the air resistance to this type of motion increases rapidly with speed break-up will occur at a faster rate as the speed of efflux is increased, whereas the varicose form of disturbance has a rate of growth independent of the speed of efflux. When a liquid jet enters the air, both types of disturbance are equally possible but at low speeds the varicose form will break up the jet first. Since the growth coefficient of the "optimum" wave-length is constant, the continuous length of the jet will be proportional to the speed, as Smith and Moss (1917) showed, but at a critical speed the sinuous disturbance will grow at the same rate as the varicose; thereafter sinuosities will break up the jet first and the length will decrease with increasing speed.

After examining a number of photographs of high-velocity jets, both Haenlein (1931) and Ohnesorge (1937) postulated a third type of break-up in which the jet is disrupted, termed "atomization" (*zerstäubung*). Ohnesorge claims, too, critical velocities delineating regions for the three types of break-up. Littaye (1939) who has examined this question further, finds that the middle region of wave formation is somewhat indefinite, slight alterations in the conditions of the experiment being sufficient to convert this intermediate type of break-up into one of the other two.

Hence it may be concluded that at high velocities in excess of a critical value, break-up is controlled by viscous and inertia forces. Any expression deduced for this break-up should involve the characteristic factors which enter into viscous flow, i.e., a velocity, a drop diameter and the kinematic viscosity of the liquid.

(Although this work does not deal with gaseous jets or jets of liquid into the selfsame liquid, it may be remarked here that the sinuous type is the only one applicable to jets in which the interfacial tension is negligible, as, for example, in the sensitive flame.)

§ 2. THE MEASUREMENT OF DROP SIZE

Although some work of a qualitative nature has been carried out on the drops from jets delivered at high speed from narrow nozzles, in connection with engine oil injectors (Kuehn (1924), Scheubel (1927) and Schweitzer (1937)), exact measurements with proper control and measurement of the factors involved are lacking. The first part of this programme of work envisaged measurements of mean drop size, varying, in turn, speed of jet relative to air, nozzle diameter and viscosity of liquid, as being the fundamental factors likely to affect break-up in the atomization region, into which the majority of these experiments fall.

It was not possible to measure drop size immediately after break-up. Instead, the drops were allowed to fall on sheets of thin blotting paper (1 ft. diam.) at some distance. The stains, made clear by admixture of a dye in the liquid, were measured and the relation between stain-diameter and drop size established by shooting single drops of the same liquid from a pipette directly on to the paper; save for very small drops, this relationship was linear. With most liquids the relation was independent of impact velocity within the range of the investigation. Water was an exception, as large drops tended to splash on landing at high speed. As at a given jet speed the drops cover a range of sizes, it is necessary to define a "mean drop diameter". This may be done in a number of ways; the drop which occurs most frequently could be taken, or that size which is responsible for the greatest fraction of the total amount of liquid discharged. Actually it is the latter interpretation which is to be given to "mean drop size" in the results as plotted on the figures of this paper.

§ 3. MEAN DROP SIZE FROM STATIONARY AND MOVING NOZZLES

Drop-size distribution curves were obtained with a jet fired by gas pressure from a stationary nozzle on a cylinder directed vertically downwards in an enclosed tower and also from a nozzle horizontally discharging backward from an aircraft. In both cases the height of fall to the paper was about 50 ft. The jet velocity was calculated from the recorded pressure in the cylinder. In the tower experiments, overlapping of stains was prevented when the drops were insufficiently dispersed by swinging the board holding the filter paper on a bifilar suspension across the line of fire during emission. In the aircraft experiments, advantage was taken of a cross-wind to scatter the drops and many papers were spread out along and to one side of the line of flight. As it was intended that from measurements of the stains, drop sizes at break-up should be derived, it was necessary to consider the evaporation which takes place while the drops are falling. This calculation has been done in another connection (Richardson, 1946) and proves that with most liquids only the smallest drops (less than 0·25 mm. in diameter at this height) require correction for this cause. Exceptions to this statement must be made for an extremely volatile liquid like carbon tetrachloride, the uncorrected mean drop being, from this cause, too low.

The stains produced by the drops were counted in 1 mm. groups (0 to 1, 1 to 2, etc.) over a strip 3 inches wide passing through the centre of the pattern on the filter paper and hence the total number and the total volume occupied by each group calculated. When small drops predominated, the lowest group was further subdivided. The total volume occupied by all drops up to a particular size was plotted against mean diameter of the group. The curves (specimens of which are shown on figures 1 and 2) show that the distribution is symmetrical, so that the mean drop size could be taken to be that corresponding to 50 per cent of the total volume discharged.

A list of the liquids used and their physical properties follows:

Liquid	Density ρ (gm./cc.)	Kinematic viscosity ν (cm^2/sec.)	Surface tension σ (dynes/cm.)
Zinc chloride	1·76	0·10	40
Soap solution	1·00	0·020	30
Titanium tetrachloride	1·76	0·006	32
Carbon tetrachloride	1·60	0·006	25
Methylene chloride	1·33	0·004	30
Methyl salicylate	1·33	0·03	35
Chlorosulphonic acid	1·77	0·03	38
Water	1·00	0·012	73
Methyl salicylate (thickened)	1·34	7·0	25
Glycerine	1·26	10·0	64
Glycerine + 20 % water	1·21	1·0	45

The experiments with stationary nozzles indicated that the mean drop diameter d depends only on the viscosity of the liquid and on jet velocity V. In fact, d was found to be inversely proportional to V (i.e. to the relative velocity between the jet and the surrounding air). Two conical nozzles 0·15 and 0·25 inch diameter were used in addition to a 5/16-inch diameter sharp-edged orifice, and these gave identical results independent of nozzle size or shape. Some later tests with a small nozzle 0·030 inch diameter confirmed this.

As an example, figures obtained with two liquids, soap solution and glycerine, are given below:

Liquid	Jet velocity V (m./sec.)	Mean drop size d (mm.)	Vd/ν
Soap solution	25	1·10	14,000
	36	0·55	10,000
	57	0·38	11,000
	68	0·30	10,000
Glycerine	88	0·88	77
	109	0·73	79

In most of the experiments with the moving nozzles, the speed of the jet relative to the nozzle was greater than the aircraft speed, but in one instance it was less. In every case the mean drop size was found to depend only on the *relative* speed, V, of the jet to the air and on the viscosity of the liquid. Here again, d was inversely proportional to V. The nozzles used varied in diameter from 0·4 to 0·7 inch, and also gave identical results independent of nozzle size.

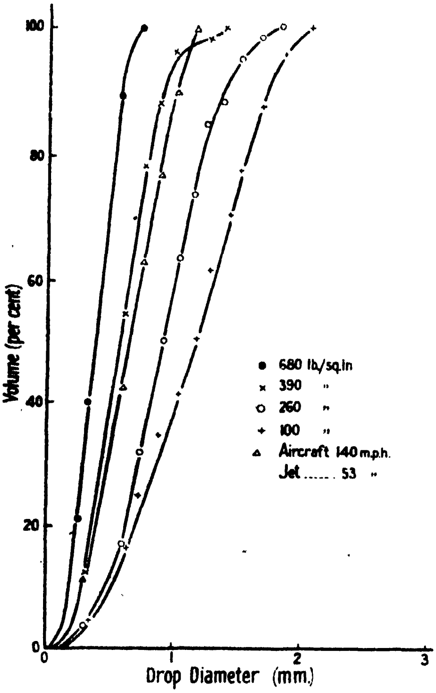


Figure 1. Zinc chloride.

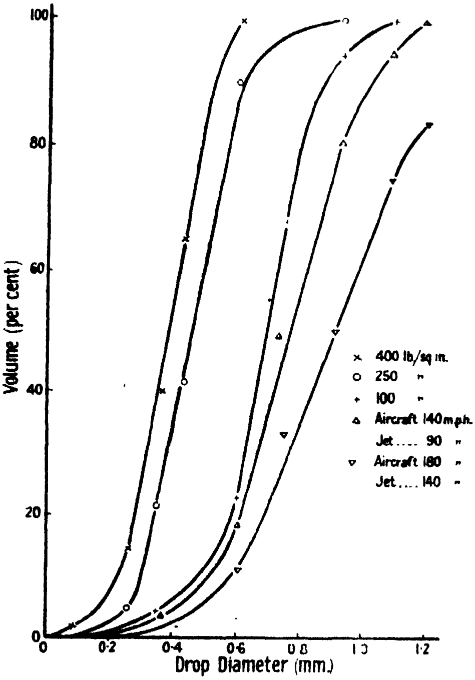


Figure 2. Titanium tetrachloride.

The surface tension of most of the liquids used is in the neighbourhood of 30 dynes per cm.; only water and glycerine have higher surface tensions, and these produced mean drop sizes expected from their respective viscosities and showed no effects which could be ascribed to surface tension. This result implies that at these speeds break-up is due to disruption of the jet.

On figure 3, Vd/ν is plotted against ν in log: log form. The results, for *both* stationary and moving nozzles, satisfy the empirical formula $Vd/\nu^{1/5} = 500$,* where V is the relative velocity between the jet and the surrounding air, d is the mean drop size, and ν is the kinematic viscosity of the liquid. Since static and moving nozzles give identical expressions for drop size, it is concluded that break-up in the "atomization" region is initiated by turbulence within the jet but is finally controlled by air friction on the jet surface. This is in agreement with the conclusion arrived at by Schweitzer (1937), who found that atomization could not be obtained without air friction, but to make air friction effective turbulence is essential.

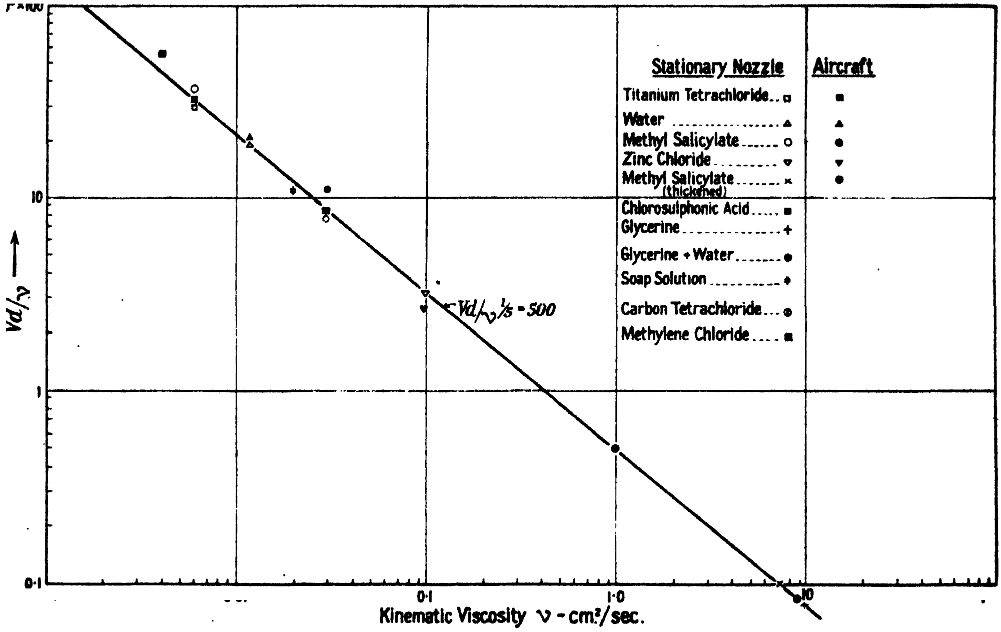


Figure 3.

§4. LIMITING DROP SIZE

It was postulated at the commencement of this paper that disruption of the jet is controlled by viscous and inertia forces and as such would predominate at high speeds. At lower speeds the varicose disturbance would be the dominant cause of break-up, while an intermediate type of break-up, in which the jet followed a sinuous motion, also occurs. We should, therefore, expect the relationship deduced from the above-mentioned experiments to cease to apply at low velocities. The varicosities then break the jet up into ovoid lumps which eventually resolve into drops comparable with the jet diameter and therefore with the nozzle diameter. Experiments (figures 4 and 5) show in fact that at low speeds the mean drop size

---

* This formula is, of course, dimensionally unsound, but could be corrected by including a term involving the kinematic viscosity of the surrounding medium. Since the latter was not varied in these experiments this term has been omitted.

The mean values of B for "thin" liquids are 17 and 12 for 50 and 125 ft. respectively. The fact that a larger drop remains entire at the smaller height can be ascribed to the shorter time over which the disrupting forces can operate. The viscous liquid—thickened methyl salicylate—shows exceptionally high values of the critical B. This indicates a superior resistance to the tendency of shearing forces to produce circulation and is paralleled by Bond and Newton's observations of viscous oil drops in water.

To summarize this part of the work: at low velocities a change in the type of break-up may limit the mean size of drop attainable from a jet, and in such a case the mean size will no longer be independent of the diameter of the nozzle. In addition, the maximum drop size will in any type of break-up be limited by the size of the largest drop which can remain unbroken under the circumstances, even at high velocities and particularly with large nozzles.

§ 5. CONTINUOUS LENGTH OF JET

The shape of the curve relating length L of the continuous portion of the jet to velocity of efflux was first described by Smith and Moss (1917). There are two significant parts of the curve, (1) a straight portion: L/D proportional to $V\sqrt{\rho D/\sigma}$, (2) a hyperbolic portion. Tyler and Richardson in further experiments (1925) showed that the latter was governed by a relation $VL/\nu = \text{constant}$. This work was done with capillary nozzles ($<0{\cdot}6$ mm. diameter).

In order to extend the data on jet length provided by the 1917 and 1925 papers, measurements of jet length as well as of mean drop-size were made with the wide nozzles used in the present work.

Both Rayleigh (1892) and Weber (1931) have obtained expressions for the optimum rate of growth of "varicosities" on a cylindrical jet of *viscous* liquid. Viscosity, of course, superposes a damping term on the growth factor. Thus Weber's expression for the damping coefficient is

$$2\mu = -\frac{3\nu}{a^2}\epsilon^2 \pm \sqrt{\left\{\frac{9\nu^2}{a^4}\epsilon^4 + \frac{2\sigma}{a^3}(\epsilon^2-\epsilon^4)\right\}},$$

where a is the jet radius and $2\pi a/\epsilon = \lambda$, the wave-length of the disturbance. If ϵ is the same for optimum rate of growth, whether the jet be viscous or not (and Rayleigh considers this to be so) $\epsilon = 1/\sqrt{2}$ and

$$2\mu_{\text{opt.}} = \frac{1}{a^2}\left(-\frac{3\nu}{2} \pm \sqrt{\frac{9}{4}\nu^2 + \frac{\sigma a}{2}}\right).$$

Taking both negative signs we obtain the following values for -2μ with the radii 0·19 cm. and 0·05 cm. of the principal nozzles used in this research:—

ν cm.^{2}/sec.	-2μ	
	a=0·19 cm.	a=0·05 cm.
0·01	65	600
0·1	70	700
1·0	140	1,930
10·00	412	6,000

Now between viscosities 0·01 and 0·1 there is a negligible change in μ and therefore in L, but for $\nu = 1$ and $\nu = 10$ the lengths are respectively twice and seven times those with negligible viscosity.

Figure 7 shows the measurements of length with a nozzle of 3·8 mm. diameter. The straight portions of the characteristic curves for varicose jets of small viscosity can be made to coincide if L/D is plotted against $V\sqrt{\rho D/\sigma}$. Since all these results are for one nozzle and there is little difference in the parameter $\sqrt{D\rho/\sigma}$ it is apparent from an inspection of the figure that this no longer holds for viscous liquids. That this effect is not peculiar to the one nozzle was shown by similar measurements on 0·5, 1 and 2 mm. nozzles (figure 8), the smallest of which overlaps

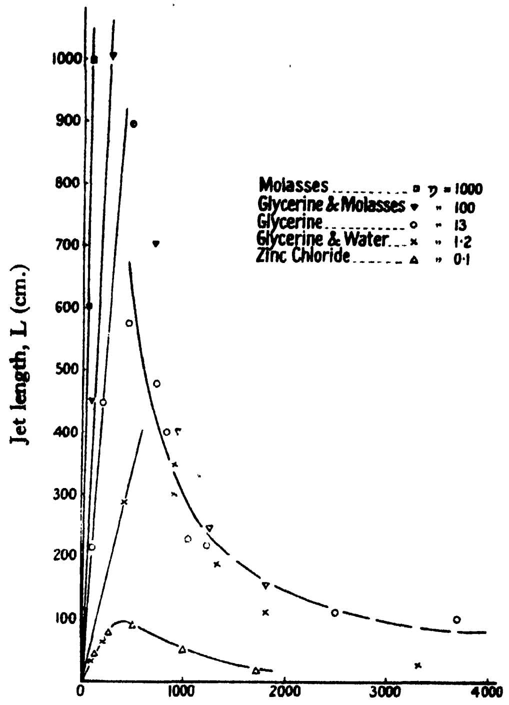


Figure 7. Results with jet 3·8 mm.

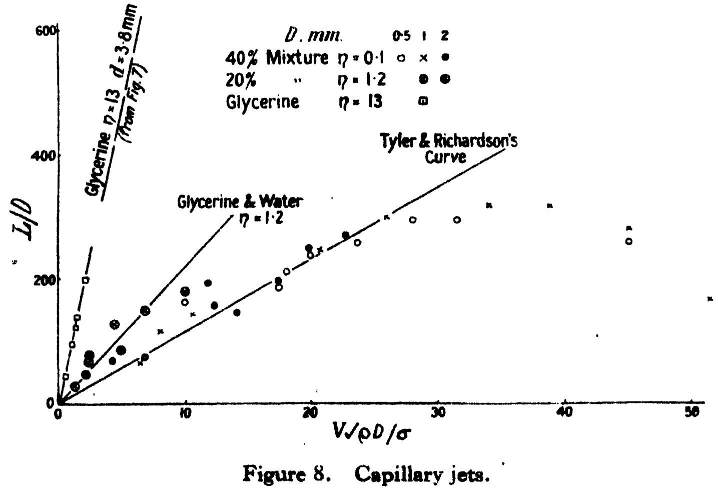


Figure 8. Capillary jets.

the range of nozzle size used by Tyler and Richardson, from whose figure 9 the dotted line on figure 8 is copied. Furthermore, the straight portion of the curves in figures 7 and 8 for $\nu = 1$ and $\nu = 10$ indicates lengths which are respectively about twice and seven times those with negligible viscosity. These figures agree well with those given above derived from Weber's theoretical expression for the damping coefficient.

Some experiments were also carried out on mercury and liquids exhibiting anomalous viscosity, one set of which, 1½% solution of rubber in petrol, is shown with the mercury results on figure 9. Pressures up to 1000 lb. per sq. in. were used on mercury. These measurements are interesting because they confirm the observation of Tyler and Richardson that on contrast to all other liquids, jet length in mercury depended to a marked degree on the shape and method of manufacture of the nozzle.

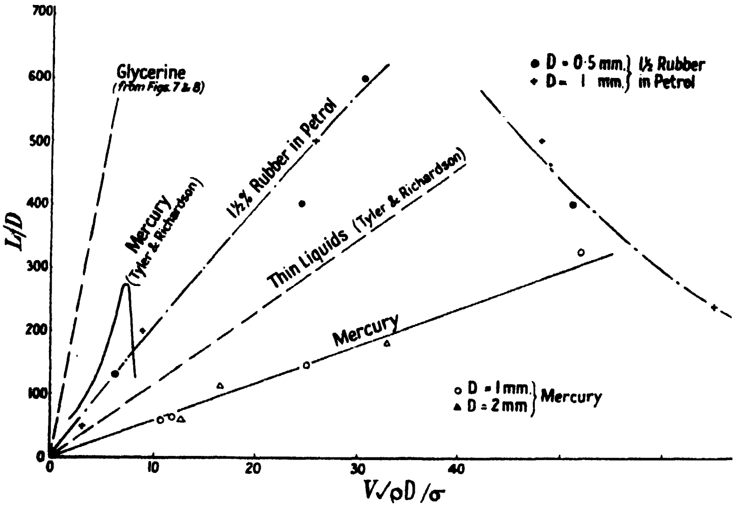


Figure 9.

The rubber solution shows a characteristic property of such fluids. Its position on figure 9 is appropriate to a viscosity of 2 or 3 poise, whereas its apparent value in an Ostwald viscometer (bore 0·5 mm.) was ½ poise. Evidently the shearing forces in the motion of the liquid through the air are comparatively light, so that its effective viscosity is large.

Some photographs were taken of the mercury jets. One of these is interesting as it shows at various sections both varicose (figure 11) and sinuous motion (figure 10) at the same time. The varicose motion eventually persists and breaks up the jet (figure 12).

ACKNOWLEDGMENTS

This work was carried out by the Armament Research Department in collaboration with the Royal Aircraft Establishment, Farnborough. The authors are indebted to the Chief Scientific Officer, Ministry of Supply, and the Director, Royal Aircraft Establishment, for permission to publish this paper.

Figure 10. Sinuous motion in a mercury jet.
V 20·6 m./sec., nozzle diameter—2 mm.

Figure 11. Varicose motion.

Figure 12. Varicose motion breaking up the jet.

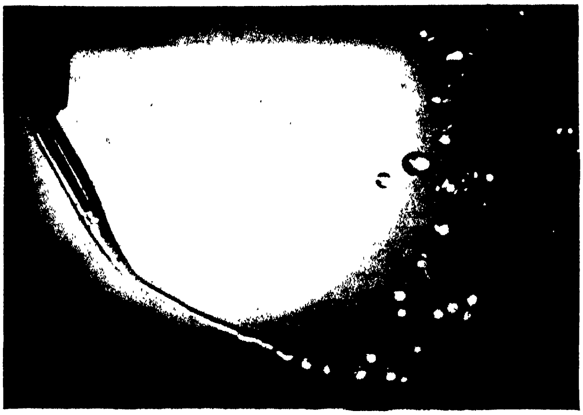


Figure 13.

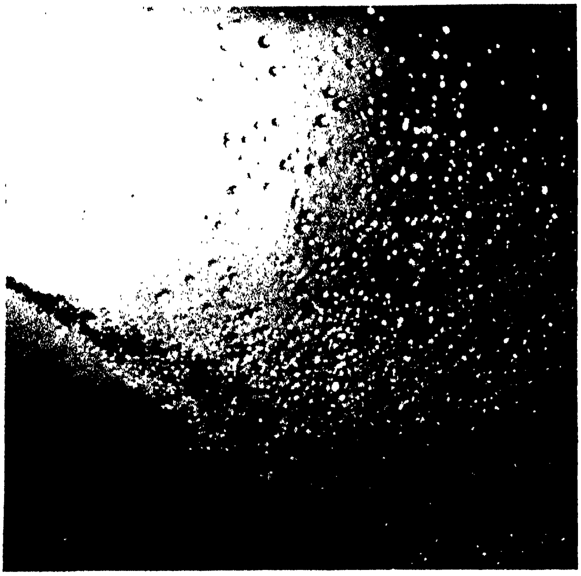


Figure 14*a*.

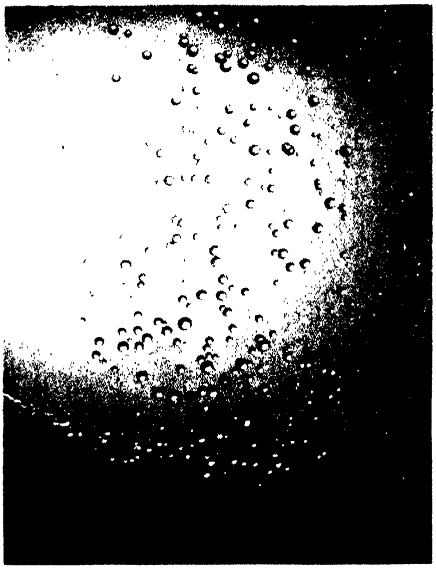


Figure 14*b*.

PART II—LIQUID INTO LIQUID (IMMISCIBLE) JETS

§ 1. INTRODUCTION

In spite of its importance as a method of manufacturing emulsions in what is known as an "atomizer", we have not been able to find any quantitative data on the break-up of liquid jets into drops by injection into water or other liquids with which they do not mix. Smith and Moss (1917) and Tyler and Watkin (1932) have extended their work on the lengths of the continuous portion of jets, noted in Part I, to cases in which one liquid is squirted from a capillary nozzle into another liquid. The same general features were observed as in the case where the liquid emerges into air, i.e., a region in which change of form of the jet without change of direction caused it to break up into drops and another at higher efflux speeds when the jet broke up at the point where the amplitude of its sinuous motion became too great for its coherence to be preserved. If the jet passed into a very viscous liquid, the growth of disturbances was hindered and the jet was longer than for the same liquid falling into air or water.

§ 2. EXPERIMENTAL WORK

The liquids forming the jets in the present research were aniline ($\nu = 0{\cdot}04$), benzene ($\nu = 0{\cdot}008$), paraffin ($\nu = 0{\cdot}023$) and various samples of fuel and lubricating oil ($\nu = 0{\cdot}12$ to $0{\cdot}68$) squirted into water, and in the case of aniline, also into salt solution, contained in a glass-sided tank 4 ft. high by 18 inches square. One per cent of an emulsifier in the water prevented early recombination of the drops. The pressure used to eject them was read on a mercury manometer and afterwards correlated with the velocity of efflux. An instantaneous photograph was taken of the field of view some distance from the nozzle. At low speeds a fast shutter was used; at high speeds a discharge (arditron) lamp. Care had to be taken in choosing a limited portion of the field for analysis that drops were not dispersed outside the depth of focus (as this caused fogging) and were not in too great a concentration (as this caused overlapping of images).

The photographs after development were projected, together with a superposed graticule, on a screen and the different sizes sorted to get summation curves similar to those of Part I. The limit of resolution was about $\frac{1}{8}$ mm. (actual size).

Such curves were obtained for aniline, paraffin and oil, and benzene into water ($\sigma/\rho = 6$, 40 and 35 c.g.s. respectively) and for aniline into brine ($\sigma/\rho = 12$) using three circular nozzles of diameters 1·0, 1·5 and 2·75 mm. at speeds up to 8 m./sec. The critical velocities (cm./sec.) for the change of régime from varicose to sinuous, according to Tyler and to Ohnesorge, are as follows:—

Liquids	Tyler's critical V			Ohnesorge's critical V		
	$D=0{\cdot}1$	$D=0{\cdot}15$	$D=0{\cdot}275$	$D=0{\cdot}1$	$D=0{\cdot}15$	$D=0{\cdot}275$
Aniline–water	46	38	28	70	50	19
Aniline–brine	65	53	40	98	70	26
Benzene–water	85	70	50	100	70	25
Paraffin–water	115	90	70	150	110	90

Comparison has been made between the Reynolds number for the mean drop diameter as observed and as calculated from figure 3 both directly and on the following basis: the abscissae of figure 3 were recalculated in terms of ν/ν_0 where ν_0 is the dynamical viscosity of the surrounding medium (i.e. 0·15 for air, to which the figure refers) and the ordinate of Vd/ν read off, corresponding to the appropriate value of ν/ν_0 for the two liquids involved. These are given in the following table:—

Liquids	ν/ν_0	Reynolds number of mean drop		
		Obs.	Calc. for ν alone	Calc. for ν/ν_0
Aniline–water	3·0	1500	7000	700
Aniline–brine	3·5	1250	5000	500
Benzene–water	0·66	5000	27000	3000
Paraffin–water	2·0	3000	10000	1200
Oil–water	10·0	550	2000	250
,,	12·5	675	1900	240
,,	25·0	450	1500	200
,,	55·0	200	600	100

The observed values of Reynolds number for the mean drop are about twice those derived on a relative-viscosity basis but much below those appropriate to the absolute viscosity of the liquid forming the jet. Evidently, when the critical velocity is exceeded, friction with the surrounding medium has a greater influence on the break-up of the jet than friction within its own substance.

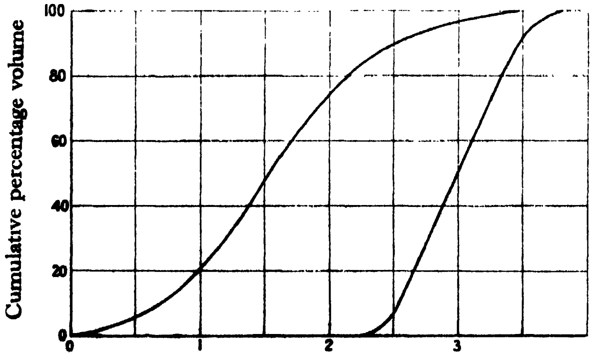


Figure 15. Drop diameter (mm.).

Below this, in the varicose régime, the drop size is governed by two additional factors, surface tension and nozzle size. The drops also tend to be more uniform in size. The difference can be seen on the specimen photographs reproduced. Figure 13 illustrates break-up in the neighbourhood of the critical velocity; figure 14*a* shows the spread of the jet in the supercritical region, while figure 14*b* illustrates a portion of the same jet sufficiently dispersed for analysis. The uniformity of size of drops at very low velocities is well seen in Tyler and Watkin's photographs of aniline–water jets, where the size is almost entirely governed by surface tension. When the speed of efflux is a negligible quantity, we have the well-known method of the "drop-weight" for the measurement of interfacial tension.

As a further comparison, two typical summation curves for drop size in the atomization (left-hand) and varicose (right-hand) régimes are shown on figure 15.

The effect of nozzle size is to limit the mean drop size at low velocities of efflux, so that if d were plotted against the reciprocal of V, we should get curves like those of figures 4 and 5. The effect on the liquid–liquid jets is shown in the following table:—

Benzene–water

	$D=0{\cdot}175$			$D=0{\cdot}15$			$D=0{\cdot}10$			
V	95	160	220	100	140	170	87	123	215	→0
d	0·50	0·25	0·25	0·30	0·35	0·25	0·25	0·20	0·20	0·5
Vd	47	40	55	30	50	42	22	25	43	—

Aniline–brine

	$D=0{\cdot}275$			$D=0{\cdot}15$			$D=0{\cdot}10$			
V	95		175	140		185	80	125	175	→0
d	0·37		0·25	0·18		0·25	0·30	0·35	0·28	0 37
Vd	35		45	24		46	24	45	50	—

In no case was break-up produced by disintegration of a drop once formed, as in certain cases of jets in air, the nozzles used being too fine to produce drops which were themselves unstable at the speeds at which they moved through the stationary liquid, so that the critical Bond number was never exceeded.

ACKNOWLEDGMENT

The work of Part II was carried out at King's College, Newcastle-upon-Tyne, by one of us (E.G.R.), who gladly acknowledges the help received from Mr. W. Bainbridge, research assistant.

REFERENCES

Bond and Newton, 1928. *Phil. Mag.*, **5**, 794.
Haenlein, 1932. *Forsch. Ing. Wes.*, **4**, 139.
Kuehn, 1924. *Der Motorwagen* (Berlin).
Littaye, 1939. *C.R. Acad. Sci., Paris*, **208**, 788 and 1705.
Ohnesorge, 1936. *Z. angew math. Mech.*, **16**, 355.
Rayleigh, 1879. *Proc. Lond. Math. Soc.*, **10**, 4.
Rayleigh, 1892. *Phil. Mag.*, **34**, 145.
Richardson, 1946. *Proc. Univ. Durham Phil. Soc.*, **10**, 394.
Scheubel, 1927. *Jahr. Wiss. Ges. Luftfahrt.*, 140.
Schweitzer, 1937. *J. Appl. Phys.*, **8**, 513.
Smith and Moss, 1917. *Proc. Roy. Soc.*, A, **93**, 373.
Tyler, 1933. *Phil. Mag.*, **16**, 504.
Tyler and Richardson, 1925. *Proc. Phys. Soc.*, **37**, 297.
Tyler and Watkin, 1932. *Phil. Mag.*, **14**, 849.
Weber, 1931. *Z. angew. math. Mech.*, **2**, 136.

THE LINES OF FORCE THROUGH NEUTRAL POINTS IN A MAGNETIC FIELD

By DAVID OWEN,
King's College, University of London

MS. received 3 *April* 1946

ABSTRACT. It is shown that, having precisely located a neutral point in a magnetic field as the point of intersection of two directional loci, it is practicable to draw the lines of force passing through the point. The theory of the field near a neutral point is treated from a geometrical point of view. Examples are shown of fields plotted when a bar magnet is set parallel or at an angle to the earth's horizontal field.

§ 1. INTRODUCTORY

In a recent paper * a method of accurately locating neutral points in a magnetic field is described, any such point being determined as the point of intersection of two directional loci which can be plotted with a small compass needle. The properties of the field near a neutral point thus assume a new practical interest. In particular, the lines of force passing through a neutral point can be plotted, starting at the known point. If a small circle be drawn with the point as centre it will be found that there are two diameters at right angles along which the compass needle will point, with one of its poles over or near the centre of the circle, showing that there are two lines of force crossing orthogonally and reversing direction at the point.

It will be shown that in general the lines of force through a neutral point consist of an isolated line in a certain direction and an infinite number of lines in the plane at right angles. If the direction of the field along the first line is towards the neutral point the directions of all the lines in the orthogonal plane are away from the point. The tangent to the first line, at the neutral point, may appropriately be termed the principal axis of the field near the neutral point. Usually the selected plane in which a field is plotted contains that axis, and in such cases, accordingly, two lines of force only will be found passing through the point in the plane of plotting.

A theoretical treatment of the field in the vicinity of a neutral point is developed from a geometrical point of view, attention being directed to lines of force and field intensities as well as to equipotential lines and surfaces.

Two examples in illustration are given of the fields plotted when a bar magnet is set horizontally in the earth's field, first with its axis parallel, then at an angle, to that field. Some interesting properties of the field due to the superposition of the field of a circular current on the earth's field are also pointed out.

---

* Owen, *Proc. Phys. Soc.*, **57**, 294, 1945.

§2. THEORY

The mathematical investigation of neutral points and lines of neutral points has usually been given with exclusive reference to electric fields.* Neutral lines, or lines of equilibrium, are of common occurrence on the surfaces of charged conductors, and whole regions of space may occur in which the potential is constant. Such phenomena have no counterpart in magnetic fields since, unlike lines of electric induction, lines of magnetic induction are always closed curves. Thus the points of emphasis in the two kinds of field are different. The fact that magnetic lines of force can be readily plotted lends a practical interest to the study of the magnetic field.

In place of the usual analytical treatment a geometrical approach is here proposed, attention being directed to lines of force and field intensities as well as to the equipotential lines and surfaces which appear to be inevitable at the outset. Take any plane containing a neutral point, and choose rectangular axes of x and y in the plane, with origin at the point. Measuring potentials from a value taken as zero at the neutral point, the magnetic potential V at any point (x,y) can be represented by the length of an ordinate erected perpendicular to the plane. A continuous "surface of potential" (not to be confused with an equipotential surface) can thus be generated which touches the plane at the origin. The curves of section of this surface by planes parallel to the x,y plane, when projected on that plane, are equipotential lines. Close to the origin all such surfaces become approximately of the second degree and, as is shown in treatises on solid geometry, belong to one of three classes: the elliptic paraboloid, the hyperbolic paraboloid, and the parabolic cylinder.

The first of these, the elliptic paraboloid, lies wholly an one side of the x,y plane, and the corresponding equipotential lines are similar ellipses.

The second, the hyperbolic paraboloid, is a saddle-back or anticlastic surface, lying partly above and partly below the x,y plane. The curves of section by planes above the reference plane are similar hyperbolas and the equipotential lines obtained by projection are of one sign of V; while the sections below the plane yield on projection the conjugate hyperbolas, with the opposite sign of V. The reference plane itself cuts the surface in a pair of straight lines through the origin, namely, the common asymptotes $V=0$ of both sets of hyperbolas.

The third class of surface, the parabolic cylinder, lies wholly on one side of the reference plane, and the sections parallel to the plane are pairs of straight lines. Such a surface would lead to a field in which all the equipotential lines were parallel, and therefore cannot represent the essentially non-uniform field near a neutral point. This consideration eliminates the parabolic cylinder as a possible form of surface of potential.

The form of the equipotential surfaces near a neutral point can now be determined. Sections by any plane through the point must, as we have shown, be either ellipses or hyperbolas. The ellipsoidal form of surface is excluded by the consideration that it would imply the existence of a field with lines of force all converging to or diverging from the neutral point, in contradiction to Gauss's theorem as applied to magnetism. The hyperbolic form of equipotential surface

---

* See Maxwell's *Electricity and Magnetism*, Vol. I, Chaps. VI and VII, and Jeans's *Electricity and Magnetism*, Chap. II.

remains as the unique general solution. The equipotential surfaces near a neutral point accordingly consist of a system of similar hyperboloids of two sheets, with potentials of one sign, and of the conjugate hyperboloids of one sheet, with potentials of the opposite sign, the two sets being separated by the asymptotic double cone $V=0$. The form of the members of such a system is indicated in figure 1.

The principal axes of the system may now be adopted as the rectangular axes of x, y and z, with axial lengths a, b and c respectively. The general equation of the whole system is

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}-\frac{z^2}{c^2}=\pm 1, \qquad \ldots\ldots(1)$$

the positive sign on the right-hand side applying to the 2-sheet hyperboloids, the negative sign to the 1-sheet hyperboloids. As a, b and c change values from zero upwards, their ratios remaining constant, the successive equipotential surfaces are described.

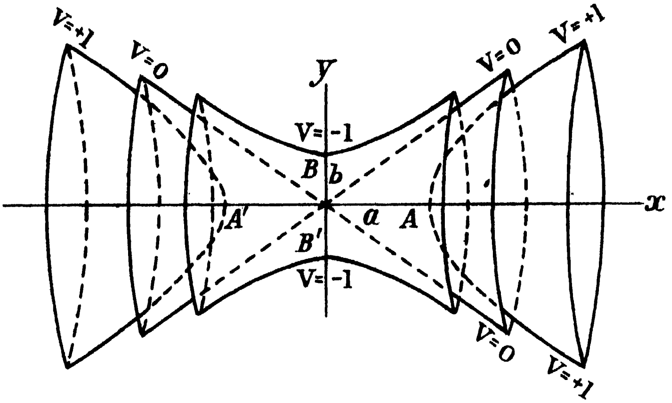


Figure 1. Equipotential surfaces close to a magnetic neutral point, showing one 2-sheet hyperboloid ($V=1$) and one 1-sheet hyperboloid ($V=-1$), separated by the asymptotic cone ($V=0$). The axes of x and y (lengths $OA=a$ and $OB=b$) are taken in the plane of the paper. The axis of z (length c) is not shown.

It can at once be seen that the lines of force through a neutral point, being orthogonals to the equipotential surfaces, consists of one line along the x or a axis, and an infinite number of lines in the perpendicular plane. The lines of force in that plane can be shown to be curves satisfying the equation $z=Ky^n$, where $n=b^2/c^2$ and K takes all values from zero to infinity.

The x-axis can thus be regarded as the principal magnetic axis of the field at the neutral point. Lines of force plotted in any plane containing that axis (as in the two practical examples given below) will show only two lines of force through the neutral point, crossing at right angles at the point; if the direction of the field is towards the neutral point along one line, it is away from it along the other.

Since the surface of potential with respect to any plane through a neutral point is of the second degree and touches the plane at that point, the potentials at the extremities of the axes of any equipotential hyperboloid may be denoted by Aa^2, $-Bb^2$ and $-Cc^2$, where A, B and C are positive constants. Thus

$\partial^2V/\partial x^2=2A$, $\partial^2V/\partial y^2=-2B$, and $\partial^2V/\partial z^2=-2C$. Applying Laplace's equation

$$\partial^2V/\partial x^2+\partial^2V/\partial y^2+\partial^2V/\partial z^2=0$$

it follows that $A-B-C=0$, thus imposing a relation between a, b and c, namely

$$1/a^2-1/b^2-1/c^2=0. \qquad \ldots\ldots(2)$$

In general, therefore, both b and c must be greater than a, so that the field gradient is steepest along the x-axis. Two special cases occur: first when $b=c=a\sqrt{2}$; next when $a=b$, $c=\infty$. In the first case the hyperboloids become surfaces of revolution about the x-axis. The two equipotential lines $V=0$ in any plane through the x-axis make angles $\tan^{-1}\sqrt{2}$ with the axis; and the field-gradient along that axis is twice that along the y- and z-axes, and of the opposite sign. This case is illustrated in figure 2. In the second case the equipotential surfaces become hyperbolic cylinders, the equipotential lines $V=0$ in the plane $z=0$ make angles of 45° with the x- and y-axes, the field-gradients are equal but of opposite sign along those axes, while the z-axis is tangent to a circle of neutral points. This case is illustrated in figure 3.

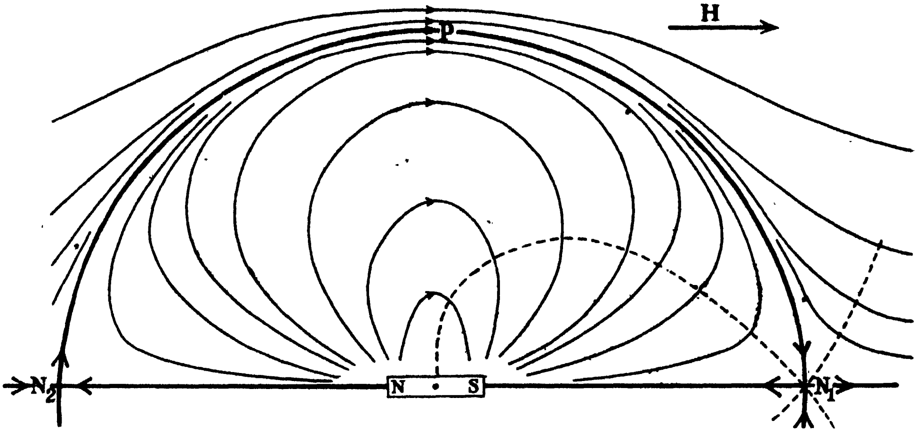


Figure 2. Field (one half, to half-scale) of a cobalt-steel magnet 1″ long, ¼″ in diameter, axis parallel to H, S-pole to the north. The lines of force through the neutral points are drawn of double thickness. The dotted lines are the plotted equipotentials through one of the neutral points, and make angles with the axis agreeing well with the theoretical value $\tan^{-1}\sqrt{2}$.

§ 3. EXPERIMENTAL

The two practical examples of plottings of fields containing neutral points relate to a simple bar magnet set with its axis horizontal in the earth's field. As the vertical component of that field has no influence on plotting in a horizontal plane, the lines of force obtained are of course identical with those due to the superposition of a uniform field H on the field of the magnet. The pair of lines of force crossing orthogonally at the neutral point are drawn of double thickness.

In the first example (figure 2) the magnet is set with its axis parallel to H, with S-pole to the north. The principal line of force, through both neutral points, is along the axis of the magnet. The second line N_1PN_2 also contains both neutral points; with the complementary half (not shown) it forms a nearly circular closed curve, a little flattened in the direction of the equatorial line of the magnet. It is interesting to note that were the magnet replaced by a doublet of equal magnetic moment this line would become a circle; and that at any point on the circle, taking the doublet as origin and measuring θ from the axis, the field intensity would be $\frac{3}{2} H \sin\theta$.

The second example illustrates a more general case, the bar magnet being set with axis perpendicular to H.

Interesting examples, easy to plot experimentally, also occur when a circular current is set with plane vertical and axis parallel to H. If the two fields oppose along the axis the pair of neutral points coincide at the centre of the coil for a certain value of the current. For current values above this the neutral points separate along the axis, while for values below it they separate along the diameter. On reversing the current the neutral points are always on a diameter and outside the coil.

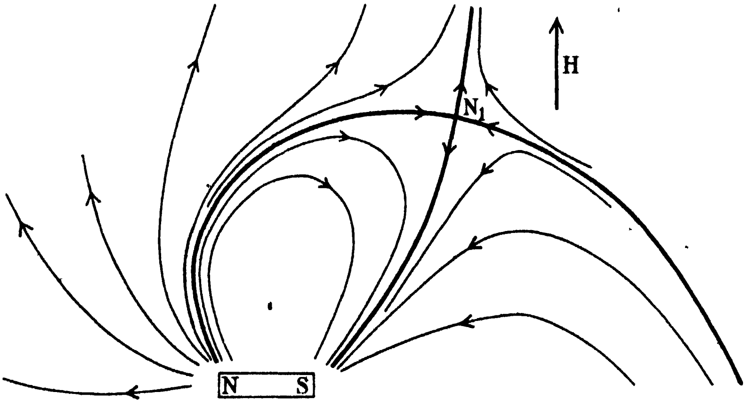


Figure 3. Upper half of field (to half-scale) due to the same cobalt-steel magnet placed with its axis perpendicular to H, N-pole to the west.

When the neutral points fall on the axis the same pair of lines of force pass through both neutral points, one coinciding with the axis, the other embracing the coil; and the pair of equipotential lines $V = 0$ make angles $\tan^{-1}\sqrt{2}$ with the axis.

When, however, the neutral points lie on a diameter a looped line of force passes through each neutral point, cutting itself orthogonally there. One of the equipotential lines through either neutral point is along a diameter, the second is parallel to the axis, so that both lines make angles of 45° with the lines of force through the point.

SENSITIVITY AND IMPEDANCE OF ELECTRO-ACOUSTIC TRANSDUCERS

By P. VIGOUREUX,

Torpedo Experimental Establishment, Greenock

MS. received 29 *May* 1946

ABSTRACT. The equivalent electric circuits of piezoelectric and magnetostriction electro-acoustic transducers are derived in terms of their mechanical constants, and the potential difference across the output terminals of a transducer placed in a given sound field is calculated in terms of the pressure of the undisturbed field and the radiation resistance of the transducer.

It is shown that, with certain reservations, the radiation resistance can itself be estimated from purely electrical measurements of the impedance of the transducer in air and in water over a range of frequencies near resonance, or even at a few selected frequencies.

The construction of impedance and admittance diagrams is explained, and it is shown how the frequency of maximum acoustic output of transducers can be obtained from electrical measurements without any acoustic measurements.

---

§ 1. INTRODUCTION

The measurement of the intensity of supersonic fields in various media, e.g. in water, is usually effected by means of a calibrated hydrophone, i.e. an electro-acoustic transducer which gives a voltage proportional to the pressure or particle velocity of the field in which it is placed. It is not easy to make a hydrophone having the same sensitivity over a wide range of supersonic frequencies, but there are methods of obtaining the calibration over the range.

In many cases, however, the measurement of the supersonic field is required only in connection with an installation comprising an electro-acoustic transducer of a different kind, used primarily, say, for under-water signalling or for submarine detection, and, with certain limitations, it is possible to make use of the same transducer for measurement of the sound field, thus avoiding the necessity for an auxiliary hydrophone. This transducer operates only at or near its resonance frequency and its vibrating face has dimensions large compared with the wavelength; it is suitable only for measurement of supersonic fields of frequencies near its operating frequency, but these are also in general the only fields of interest for the practical use of the transducer. The only data required for this measurement are the face area and the radiation resistance of the transducer, and the only observation required is that of the potential difference which appears at the terminals of the transducer as a result of the sound field in which it is placed.

The radiation resistance can itself be derived from purely electrical measurements of the impedance of the transducer in air and water over a range of frequencies near resonance, or even at a few selected frequencies, provided it can be ascertained that the mode of vibration is the same in air as in water.

The theory underlying those measurements is given below for piezoelectric and magnetostrictive transducers.

§2. MECHANICAL IMPEDANCE OF TRANSDUCER

The only accessible parts of the transducer, as far as the measurements considered here are concerned, are its vibrating face and its two terminals A and B. For more generality let an external impedance Z_e be connected across A, B.

Denote by $\mathscr{Z}_T$ the mechanical impedance of the transducer under these conditions, i.e. the quotient of the force F applied to the face of the transducer and the resulting velocity u of the face.

When the transducer is immersed in an elastic medium, displacement of the face is resisted not only by the impedance $\mathscr{Z}_T$ of the transducer, but also by an additional mechanical impedance $\mathscr{Z}_W$ which depends on the area A of the face and on the frequency; the velocity of the face is, therefore, inversely proportional to $\mathscr{Z}_W+\mathscr{Z}_T$. It must also be proportional to the free field pressure p, and is, therefore, equal to $mp/(\mathscr{Z}_W+\mathscr{Z}_T)$, where m is independent of $\mathscr{Z}_T$ but may depend on the face area and on frequency. But when $\mathscr{Z}_T$ is equal to $\mathscr{Z}_W$, the sound field in front of the face is not disturbed by the transducer,* for the impedance would not be altered by it; accordingly, in that case, the velocity of the face is the same as the free field particle velocity, viz. $p/\rho c$; hence

$$\frac{mp}{2\mathscr{Z}_W}=\frac{p}{\rho c} \quad \text{or} \quad m=\frac{2\mathscr{Z}_W}{\rho c}.$$

The velocity can therefore be written

$$u=\frac{2p\mathscr{Z}_W}{\rho c(\mathscr{Z}_W+\mathscr{Z}_T)}.$$

In the case when the dimensions of the transducer face are large compared with the wave-length, the "plane wave damping" condition is approached, and the impedance $\mathscr{Z}_W$ is nearly equal to the pure resistance ρcA. In this case the above formula reduces to

$$u=\frac{2pA}{\rho cA+\mathscr{Z}_T}. \qquad \text{......(1)}$$

Expressions must now be obtained for the impedance $\mathscr{Z}_T$ for various types of transducers. These types can be divided into two classes, one of which includes the electrostatic and piezoelectric types, the other the electromagnetic and magnetostriction types.

Let the electrical impedance from A to B, i.e. the quotient of voltage applied to AB, and current in AB when the face of the transducer is clamped, be Z, and let the mechanical impedance, i.e. the quotient of force applied to the face and velocity of the face when the terminals A, B are on open circuit, be $\mathscr{Z}$. In the absence of electrical input, if a force F applied to the face produces a velocity u, the total power input is the product $F.u$, and when electromechanical transfer of energy is allowed to occur, this power must be equal to the sum of the powers dissipated mechanically and electrically.

In the piezoelectric case the quantity of electricity liberated at the electrodes is proportional to face displacement, in other words, face velocity u produces in the

---

* Actually there is the additional requirement that the acoustic impedance per unit area be uniform over the face of the transducer, otherwise the field is distorted. This additional condition is approached when the face dimensions are large compared with the wave-length. The absence of a rim and other stationary supports near the face is also implied, as they would cause distortion.

impedance Z_p formed by Z and Z_e in parallel a current ku, where k is a constant of the transducer independent of frequency. Since the electrical power is $Z_p i \,.\, i$, the power equation is

$$F \,.\, u = \mathscr{Z} u \,.\, u + Z_p ku \,.\, ku,$$

from which, after division by u^2, the total impedance is found to be

$$\mathscr{Z}_{\mathrm{T}} = \frac{F}{u} = \mathscr{Z} + k^2 Z_p. \qquad \ldots\ldots(2)$$

In the magnetostriction case, on the other hand, the magnetic flux linking the winding is proportional to face displacement; in other words, velocity u produces in the loop of total impedance Z_s formed by one or more parts of Z (e.g. inductance and series resistance to the winding), in series with Z_e and the remaining parts (e.g. shunt resistance of the winding), an e.m.f. hu where h is a constant of the transducer independent of frequency. Since the electrical power is $e \,.\, e/Z_s$, the power equation is

$$F \,.\, u = \mathscr{Z} u \,.\, u + hu \,.\, hu/Z_s,$$

from which the total impedance is found to be

$$\mathscr{Z}_{\mathrm{T}} = \frac{F}{u} = \mathscr{Z} + h^2/Z_s. \qquad \ldots\ldots(3)$$

§ 3. SENSITIVITY OF PIEZOELECTRIC RECEIVER

In most cases the quantity observed or used is not the face velocity but the electric potential v to which it gives rise across AB. In the piezoelectric case this voltage is equal to the current through Z_p multipled by Z_p. The current is equal to ku, and by insertion of the value of u from (1), substitution of the value of $\mathscr{Z}_{\mathrm{T}}$ from (2) and abbreviation of the motional impedance $(\rho cA + \mathscr{Z})/k^2$ by J, it is found that

$$\begin{aligned} i = ku &= \frac{2pkA}{\rho cA + \mathscr{Z}_{\mathrm{T}}} \\ &= \frac{2pkA}{\rho cA + \mathscr{Z} + k^2 Z_p} \\ &= \frac{2pA}{k} \cdot \frac{1}{Z_p + J}, \end{aligned} \qquad \ldots\ldots(4)$$

from which it appears that the representation of the piezoelectric receiver can be completed by addition to Z_p of an electric curcuit J in which there is an e.m.f. e equal to $2pA/k$. If a, b, g are the mass, damping and restoring constants of the transducer, $\mathscr{Z}$ is equal to $aD + b + g/D$, where D is the time differential operator. If this value of $\mathscr{Z}$ be inserted in the formula for J, it is found that J is equal to $ND + S_i + 1/KD + \rho cA/k^2$, where N, S_i and K are written for a/k^2, b/k^2 and k^2/g respectively, and represent the equivalent series inductance, internal resistance and capacitance of the transducer. By analogy the term $\rho cA/k^2$ must be identified with the radiation resistance S of the transducer; this relation gives

$$k = \sqrt{\frac{\rho cA}{S}}, \qquad \ldots\ldots(5)$$

and, therefore,
$$e=\frac{2pA}{k}=2p\sqrt{\frac{AS}{\rho c}}, \qquad \ldots\ldots(6)$$

and the circuit of figure 1, into which is injected an e.m.f. e independent of frequency, gives an electrical representation of the behaviour of the receiver. In that circuit, the impedance Z_p comprises the capacitance c of the transducer and the external impedance Z_e in parallel.

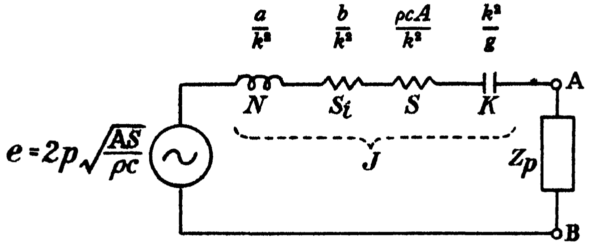


Figure 1. Electrical equivalent of piezoelectric receiver.

§4. SENSITIVITY OF MAGNETOSTRICTION RECEIVER

In the magnetostriction case, the voltage v across AB is equal to that produced by an e.m.f. hu in the winding L which forms part of the series circuit Z_s. This e.m.f. hu produces a current i in the winding and the rest of Z_s, and the terminal voltage is obtained by multiplication of i by the lumped impedance of all elements between A and B excluding that containing L. Insertion of the value of u from (1), substitution of the value of $\mathscr{Z}_{\mathrm{T}}$ from (3), and abbreviation of the motional admittance $(\rho cA+\mathscr{Z})/h^2$ by G yields

$$\begin{aligned} i=\frac{hu}{Z_s}&=\frac{2phA}{(\rho cA+\mathscr{Z}_{\mathrm{T}})Z_s}\\ &=\frac{2phA}{\left(\rho cA+\mathscr{Z}+\dfrac{h^2}{Z_s}\right)Z_s}\\ &=\frac{2pA}{h}\cdot\frac{1}{1+GZ_s}, \end{aligned} \qquad \ldots\ldots(7)$$

from which it appears that the representation of the magnetostriction receiver can be completed by addition across Z_s of an admittance G, the whole being fed with a current i_0 equal to $2pA/h$. Since as before the mechanical impedance $\mathscr{Z}$ may be represented by $aD+b+g/D$, where D is the time differential operator, the admittance G is equal to $KD+\frac{1}{S_i}+\frac{1}{ND}+\frac{\rho cA}{h^2}$, where K, S_i and N are equal to a/h^2, h^2/b and h^2/g respectively, and represent the equivalent parallel capacitance, resistance and inductance of the transducer. By analogy, the resistance $h^2/\rho cA$ must be identified with the radiation resistance S of the transducer. This relation gives

$$h=\sqrt{\rho cAS} \qquad \ldots\ldots(8)$$

and
$$i_0=\frac{2pA}{h}=2p\sqrt{\frac{A}{\rho cS}}, \qquad \ldots\ldots(9)$$

and the circuit of figure 2 fed with a current i_0 independent of frequency gives an electrical representation of the behaviour of the receiver.

In general the impedance Z consists of the inductance L of the winding and its resistive component. As the loss in the winding is chiefly due to eddy currents, a shunt resistance is a better representation of it than a series resistance, because the former gives a phase angle proportional to frequency as required by eddy-current loss. This shunt resistance r must be inserted across $L+G$ in the electrical representation. Consideration of the case when Z_e is infinite shows that it must not be inserted across L alone, because it would then have no influence on the terminal voltage, which in reality it tends to reduce, since it represents loss by eddy currents.

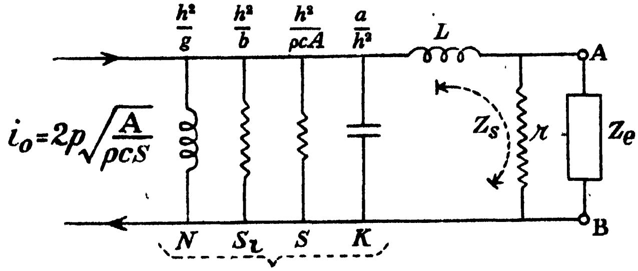
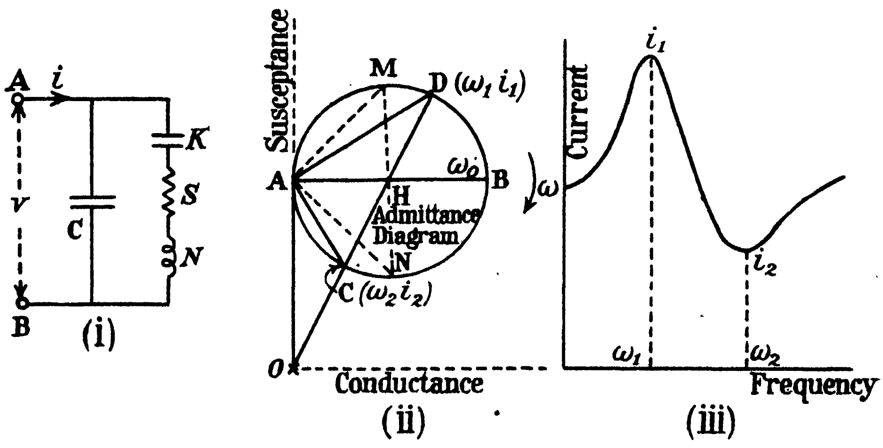


Figure 2. Electrical equivalent of magnetostriction receiver.

§5. IMPEDANCE OF PIEZOELECTRIC TRANSDUCER

With the reservations made in § 1, the radiation resistance S can be determined by purely electrical measurements.

In the case of the piezoelectric transducer, which is a fairly high-impedance device, there is no difficulty in maintaining a constant voltage across the terminals. The current observed is then proportional to admittance. In the absence of an external impedance Z_e, the impedance Z_p of figure 1 reduces to that of the condenser C, as in figure 3 (i), and by measuring admittance at frequencies well above



$$\omega_0=\sqrt{\omega_1\omega_2} \text{ approx.}, \quad S=\frac{v}{i_1-i_2}, \quad C=\frac{\sqrt{i_1 i_2}}{v\omega_0},$$

$$Q=\frac{1}{SK\omega_0}=\frac{\sqrt{\omega_1\omega_2}}{\omega_2-\omega_1}\cdot\frac{1}{2}\left(\sqrt{\frac{i_1}{i_2}}+\sqrt{\frac{i_2}{i_1}}\right) \text{ approx.}$$

Figure 3. Analysis of piezoelectric transducer.

and well below resonance, the admittance of the condenser alone can be evaluated at any frequency by proportion, and subtracted vectorially from the admittance of the whole network. If the remaining admittance is plotted vectorially, a circle of diameter $1/(S+S_i)$ is obtained, and the Q value $1/(S+S_i)K\omega_0$ is equal to $\omega_0/\Delta\omega$ where $\Delta\omega$ is the frequency difference between the two quadrantal positions, i.e. the two positions M and N, figure 3 (ii), for which the vectors are at 45° to the diameter. The constants of the transducer are thus completely determined.

In order to separate S from S_i, it is necessary to perform measurements with the transducer in air and in water or other medium in which it is used. The value of ρc for air is so much smaller than it is for water that in many cases it can be neglected altogether, but anyhow a correction can be applied. If the linearity of the transducer has not previously been established, the measurement in air must be effected with a voltage which gives the same amplitude of vibration as in water; in other words, the difference between the maximum and minimum currents must be the same in each case. Even then, subsidiary vibrations, which occur more readily in air than in water, may vitiate the results.

When the Q value is high and speed of measurement is desired rather than great accuracy, it is sufficient to observe only the two maximum and minimum currents i_1, i_2, for an applied voltage v, and the corresponding frequencies ω_1, ω_2 (figure 3 (iii)). It is shown in the Appendix that the following approximations hold:

$$\omega_0 \simeq \sqrt{\omega_1\omega_2}, \qquad \text{......(10)}$$

$$S = \frac{v}{i_1 - i_2}, \qquad \text{......(11)}$$

where, for brevity, S is written for the sum of the internal and radiation resistances,

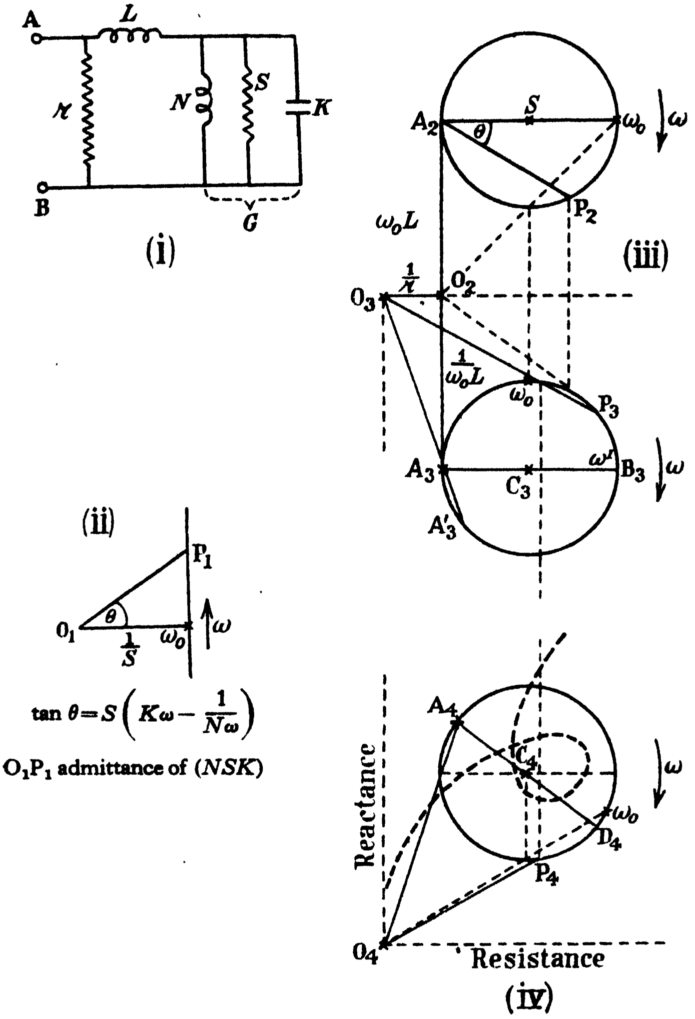
$$C = \frac{\sqrt{i_1 i_2}}{v\omega_0}, \qquad \text{......(12)}$$

$$Q = \frac{1}{SK\omega_0} = \frac{\sqrt{\omega_1\omega_2}}{\omega_2 - \omega_1} \cdot \frac{1}{2}\left(\sqrt{\frac{i_1}{i_2}} + \sqrt{\frac{i_2}{i_1}}\right). \qquad \text{......(13)}$$

The analysis of the piezoelectric transducer into its components is thereby complete. It must, however, be emphasized that this method, whilst valuable for its rapidity, is not comparable in accuracy with the more elaborate measurement of admittance at a number of frequencies and plot of the circle diagram, especially when the Q value is low.

§6. IMPEDANCE OF MAGNETOSTRICTION TRANSDUCER

As the magnetostriction transducer is a low-impedance device, a convenient way of measuring its electrical performance is to send a known current through it and to measure the voltage at the terminals for a number of frequencies. The impedance in series with L in figure 2 reduces in this case to the shunt resistance r of the winding L as in figure 4 (i). This measurement gives the impedance of the whole device. To obtain the motional impedance it is necessary to subtract the impedance of the winding. This impedance can be evaluated at any frequency by proportion if readings at two frequencies, one well below, the other well above resonance, are taken. The ends of the motional impedance vectors should lie on a circle of diameter approximately equal to S, where for brevity S is taken as the resultant of the internal and radiation resistances of figure 2. The Q value, which in this case is the resistance multiplied by $K\omega_0$, is equal to $\omega_0/\Delta\omega$, where $\Delta\omega$ is the difference between the quadrantal frequencies. The constants of the transducer are thus completely determined, and if the measurements are taken in air and in water, the radiation resistance can be separated from the frictional resistance, subject to limitations of a nature similar to those mentioned in the piezoelectric case.



$\tan\theta = S\left(K\omega - \frac{1}{N\omega}\right)$

O_1P_1 admittance of (NSK)

O_2P_2 impedance of
$L+(NSK)$

O_3P_3 admittance of
$L+(NSK)$ and r

$A_3B_3 = \frac{S}{\omega_0^2L^2}$

O_4P_4 impedance of
$L+(NSK)$ and r

$$\begin{aligned} A_4D_4 &= A_3B_3\frac{O_4A_4}{O_3A_3'} \\ &= \frac{A_3B_3}{O_3A_3\cdot O_3A_3'} \\ &= \frac{A_3B_3}{O_3C_3^2-A_3C_3^2} \\ &= \frac{S}{1+\frac{S}{r}+\frac{\omega_0^2L^2}{r^2}} \\ &= S \text{ approximately} \end{aligned}$$

Figure 4. Impedance diagram for magnetostriction transducer.

Even when the Q value is large, as for instance in the measurement in air, the quick method described for piezoelectric transducers, whereby all the constants can be determined from the two maxima and the corresponding frequencies, does not prove so convenient here, because the loss in the winding cannot be neglected as was the loss in the condenser, and in consequence it cannot be assumed that the circle is tangential to the impedance vector of the winding. The formulae of

§ 5 and figure 3 come out easily and simply, precisely because OA is a tangent to the circle.

For completeness, the ideal diagram of the magnetostriction transducer is given in figure 4, but in practice the reactance ωL of the winding varies considerably over the range of frequencies concerned, and cannot be regarded as constant as in the diagram. As a result of this variation the diagram is distorted into a loop, roughly indicated by the dotted curve of figure 4 (iv), and it is only after the subtraction of impedances mentioned above has been performed that a proper impedance circle is obtained.

Even after subtraction, the admittance motional circle is not always tangential to O_2A_3, as in figure 4 (iii), which is obtained under the assumption that the whole of the winding loss is due to eddy currents and, therefore, equivalent to a shunt resistance r. Allowance of a series resistance leads to a tilted admittance circle as well as a tilted impedance circle.

§ 7. ALTERNATIVE EQUIVALENT CIRCUIT

It is possible to obtain an almost identical admittance on the assumption that the motional impedance of the magnetostriction transducer is a series circuit N', S', K' connected in parallel with L and with r, as in figure 5 (i). In order to examine the extent to which the two circuits are equivalent, it is sufficient to calculate and inspect the expressions for their admittances Y, Y' for any frequency ω; since r occurs alike in both, it can be left out. Then

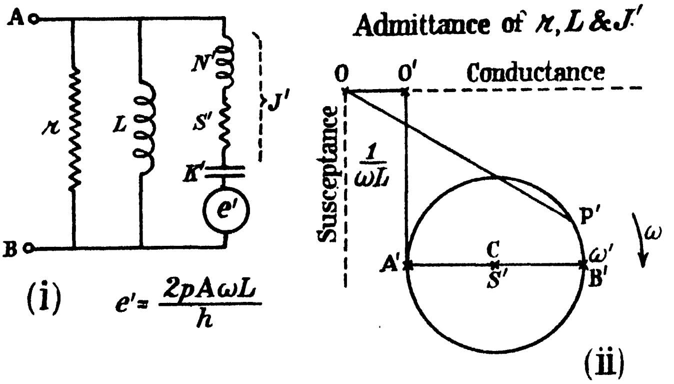


Figure 5. Alternative electrical equivalent of magnetostriction receiver.

$$\left.\begin{aligned} Y &= \frac{1}{j\omega L}\cdot\frac{1-\omega^2NK+j\omega N/S}{1+\dfrac{N}{L}-\omega^2NK+j\omega N/S}, \\ Y' &= \frac{1}{j\omega L}\cdot\frac{1-\omega^2N'K'\left(1+\dfrac{L}{N'}\right)+j\omega K'S'}{1-\omega^2N'K'+j\omega K'S'}. \end{aligned}\right\} \qquad \ldots\ldots(14)$$

It is not possible to choose N', S', K' so as to make Y' equal to Y at all frequencies; but in the neighbourhood of resonance the two expressions will be very nearly equal if

$$NN' = L^2, \quad SS' = \frac{L^2}{KN}\left(1+\frac{N}{L}\right), \quad \text{and} \quad \frac{K'}{K} = \frac{N^2}{L^2}\frac{1}{1+N/L}. \qquad \ldots\ldots(15)$$

In particular, these values give identical admittance at the frequency for which the motional admittance is a maximum. This frequency, corresponding to B_3 in figure 4 (iii), and B' in figure 5 (ii), is given by

$$\omega^2 NK = 1 + \frac{N}{L}, \quad \text{or} \quad \omega^2 = \omega_0^2\left(1 + \frac{N}{L}\right)$$

and the motional admittance is $1/S'$.

Apart, however, from the inability of the series circuit exactly to reproduce the behaviour of the transducer at all frequencies, its use as a receiver equivalent presents another difficulty: comparison of the two circuits shows that for equality of electrical output from a given sound field, an e.m.f. e' equal to $2pA\omega L/h$ must be supposed injected into the series equivalent circuit. As this e.m.f. is not independent of frequency, the series representation, although sometimes convenient, is not strictly correct for the magnetostriction case.

§ 8. FREQUENCY OF MAXIMUM RADIATION

When the transducer is excited electrically, it is often the frequency of maximum radiation which it is desired to find, rather than that of maximum or minimum impedance. In general this frequency depends on whether the transducer is excited from constant voltage or fed with constant current, but it can always be determined by purely electrical measurements. It has been pointed out that as far as impedance or admittance diagrams are concerned, it is immaterial whether the motional component of the transducer be represented by a series network of impedance J, or a parallel network of admittance G.

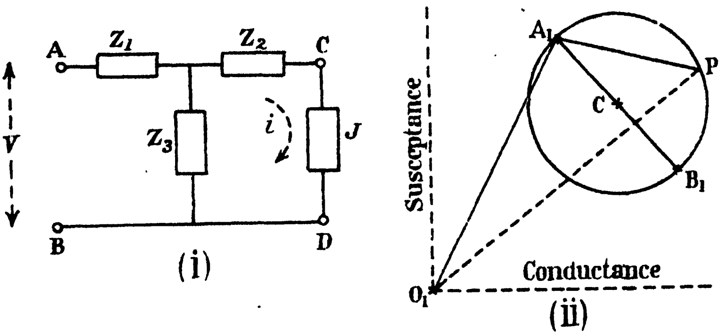


Figure 6. Admittance diagram of an electro-acoustic transducer.

To study the constant voltage case, it is convenient to use the series equivalent J. Now the whole of the purely electrical network between A and B, figure 1 or figure 5(i), can be replaced by a star network Z_1, Z_2, Z_3, connected to A, B at one end, and to J at the other, as in figure 6 (i). With constant voltage applied to AB, the current or admittance O_1P into A, B will be the resultant of two vectors, the slowly varying current O_1A_1 in Z_3, figure 6 (ii), and the rapidly varying current A_1P in J. If the admittance diagram is loop-shaped, the motional admittance circle can always be derived by subtraction of the slowly varying components, as explained in §§ 5 and 6; the origin A_1 of the circle, which is a point on the circumference, is located in the process. If the diameter A_1B_1 be drawn, the extremity B_1

determines the frequency at which the current i in J is a maximum. Since the radiated power is Si^2, this frequency is also the frequency of maximum radiation for constant applied voltage.

The constant-current case can be studied by transforming the purely electrical network between A, B, figure 2, into a Π network Z_1, Z_2, Z_3 connected to A, B at one end, and to the parallel equivalent admittance G at the other, as in figure 7 (i). With constant current fed into A, B the voltage or impedance O_2P across AB will be the resultant of two vectors, the slowly varying voltage O_2A_2 across Z_2, figure 7 (ii), and the rapidly varying voltage A_2P across G. If the impedance diagram is loop-shaped, the motional impedance circle can again be derived by subtraction of the slowly varying components, and the extremity B_2 of the diameter through the origin A_2 determines the frequency for which the voltage v across G is greatest. Since the radiated power is v^2/S, this frequency is also the frequency of maximum radiation for constant current fed to the transducer.

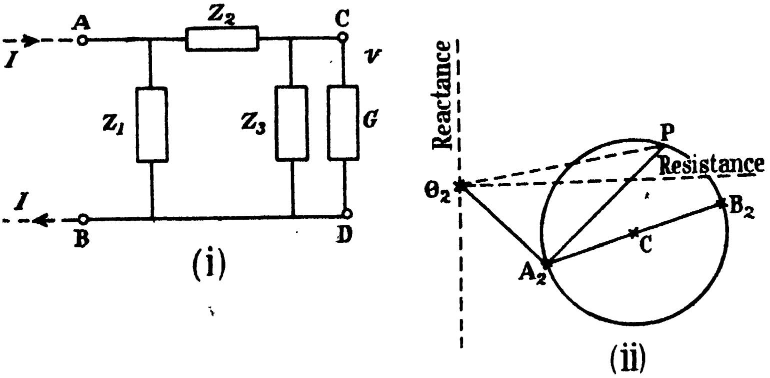


Figure 7. Impedance diagram of an electro-acoustic transducer.

In general, a transducer is neither excited from constant voltage nor fed with constant current, but is coupled to a generator of constant e.m.f. E, having a certain internal impedance, say Z_e. But the complete system can be represented by a constant voltage E applied to the impedance of the transducer and the impedance Z_e in series. Thus this general case reduces to that of constant applied voltage, and can be solved by production of the admittance diagram, figure 6.

In practice, condensers or coils are added to the circuit so as to "tune" it, i.e. to arrange that its admittance at the frequency of maximum radiation would be a maximum if the transducer were clamped. The condensers or coils also serve to "match" the transducer to the generator, i.e. to ensure that half the power generated is radiated. On account of the "tuning", the electrical impedances vary with frequency in the neighbourhood of resonance nearly as much as the motional impedance, and the admittance diagram cannot be derived in the simple way illustrated in figures 3 and 4. But in this case a readier method of determining the frequency of maximum radiation and at the same time of ensuring correct tuning is to plot current against frequency, when the "M" or double-hump curve characteristic of coupled circuits is obtained. Trial adjustment of the tuning condenser, without entailing much labour, secures a symmetrical curve, the minimum giving the frequency of maximum radiation.

ACKNOWLEDGMENT

Thanks are due to the Chief of the Royal Naval Scientific Service for permission to publish this paper.

APPENDIX

Rapid analysis of piezoelectric transducer

When the Q value is high, i.e. when $SK\omega_0$ is small, the ends of the admittance vectors lie almost exactly on a circle. In figure 3 (ii), AB is proportional to $1/S$, where for brevity S is written for the sum of the internal and radiation resistances, and OA is proportional to $C\omega_0$. The approximate values are

$$S=\frac{v}{\mathrm{AB}}=\frac{v}{i_1-i_2}, \qquad \ldots\ldots(16)$$

$$C=\frac{\mathrm{AO}}{v\omega_0}=\frac{\sqrt{\mathrm{OC}\,.\,\mathrm{OD}}}{v\omega_0}=\frac{\sqrt{i_1 i_2}}{v\omega_0}. \qquad \ldots\ldots(17)$$

The frequencies ω_1, ω_2 are related to the transducer constants by the formulae

$$\tan\mathrm{DAB}=\frac{\dfrac{1}{\omega_1 K}-\omega_1 N}{S}=\frac{\omega_0^2-\omega_1^2}{\omega_0\omega_1}Q, \qquad \ldots\ldots(18)$$

$$\tan\mathrm{CAB}=\frac{\omega_2 N-\dfrac{1}{\omega_2 K}}{S}=\frac{\omega_2^2-\omega_0^2}{\omega_0\omega_2}Q. \qquad \ldots\ldots(19)$$

But also

$$\tan\mathrm{DAB}=\tan\tfrac{1}{2}\mathrm{AHO}=\sqrt{\frac{1-\cos\mathrm{AHO}}{1+\cos\mathrm{AHO}}}=\sqrt{\frac{\mathrm{OH}-\mathrm{AH}}{\mathrm{OH}+\mathrm{AH}}}=\sqrt{\frac{\mathrm{OC}}{\mathrm{OD}}}=\sqrt{\frac{i_2}{i_1}}, \qquad \ldots\ldots(20)$$

$$\tan\mathrm{CAB}=\tan\tfrac{1}{2}\mathrm{AHD}=\sqrt{\frac{1-\cos\mathrm{AHD}}{1+\cos\mathrm{AHD}}}=\sqrt{\frac{\mathrm{OH}+\mathrm{AH}}{\mathrm{OH}-\mathrm{AH}}}=\sqrt{\frac{\mathrm{OD}}{\mathrm{OC}}}=\sqrt{\frac{i_1}{i_2}}. \qquad \ldots\ldots(21)$$

Insertion of these values in (18), (19) and division gives after some rearrangement

$$\omega_0^2=\omega_1\omega_2\frac{\omega_1 i_1+\omega_2 i_2}{\omega_2 i_1+\omega_1 i_2}=\omega_1\omega_2(1-\alpha) \qquad \ldots\ldots(22)$$

where

$$\alpha\equiv\frac{(\omega_2-\omega_1)(i_1-i_2)}{\omega_2 i_1+\omega_1 i_2}$$

is small.

Thus to a first approximation ω_0 is equal to $\sqrt{\omega_1\omega_2}$.

The value of Q can be found by adding equations (20) and (21) and equating to the sum of (18) and (19) which, after some reduction, gives

$$Q(\omega_0^2+\omega_1\omega_2)(\omega_2-\omega_1)=\omega_0\omega_1\omega_2\left(\sqrt{\frac{i_1}{i_2}}+\sqrt{\frac{i_2}{i_1}}\right)$$

or

$$Q=\frac{\sqrt{\frac{i_1}{i_2}}+\sqrt{\frac{i_2}{i_1}}}{\omega_2-\omega_1}\cdot\frac{\sqrt{\omega_1\omega_2}}{\frac{\omega_0}{\sqrt{\omega_1\omega_2}}+\frac{\sqrt{\omega_1\omega_2}}{\omega_0}}.$$

By (22), the denominator of the second factor becomes

$$\frac{\omega_0}{\sqrt{\omega_1\omega_2}}+\frac{\sqrt{\omega_1\omega_2}}{\omega_0}=(1-\alpha)^{\frac{1}{2}}+(1-\alpha)^{-\frac{1}{2}}$$
$$=2(1+\tfrac{1}{8}\alpha^2),$$

which is in general very nearly equal to 2. Thus the formula becomes

$$Q=\frac{\sqrt{\omega_1\omega_2}}{\omega_2-\omega_1}\cdot\frac{1}{2}\left(\sqrt{\frac{i_1}{i_2}}+\sqrt{\frac{i_2}{i_1}}\right). \qquad \text{......(23)}$$

Since Q is equal to $SK\omega_0$, K is obtained from (23) and (16).

LIGHT ABSORPTION AND SELECTIVE PHOTO-EFFECT IN ADSORBED LAYERS

By H. FRÖHLICH and R. A. SACK,*
H. H. Wills Physical Laboratory, University of Bristol

* Now at Department of Applied Mathematics, University of Liverpool

MS. received 1 *July* 1946

ABSTRACT. An estimate is often required of the absorption of light by a single layer of atoms (in particular alkali metal atoms) adsorbed on a non-metallic surface. Such an absorption is accompanied in many cases by a selective photoelectric effect at comparatively long wave-lengths. In the present paper the probability is calculated of a light quantum being absorbed by such a layer, and a mechanism suggested for the emission of the photo-electrons.

§1. ABSORPTION OF INCIDENT LIGHT

The absorption of a light quantum leads to the transition of an electron in an adsorbed atom from the ground state to a higher level. As with the absorption of light by atoms dissolved in solids, we should expect a considerable broadening of the absorption line, i.e. practically an absorption band about the frequency ν_0, where $h\nu_0$ is the energy difference between the ground level and the excited level of the electron. For reasons shown by Mott and Gurney (1940, p. 116) the width of the absorption band is proportional to $\sqrt{T}$, where T is the absolute temperature.

To find a general expression for the absorption we bear in mind that, as far as the transition from the ground state to the excited state is concerned, a number N of adsorbed atoms behave like Nf classical harmonic oscillators of electronic charge and mass, where f denotes the oscillator strength connected with the transition. Let

$$E = E_0 \cos 2\pi\nu t \qquad \ldots\ldots(1)$$

be the homogeneous electric field acting on the adsorbed layer; the polarization induced in it will be of the form

$$P = P_1 \cos 2\pi\nu t + P_2 \sin 2\pi\nu t. \qquad \ldots\ldots(2)$$

It is useful to define a complex dielectric constant

$$\epsilon = \epsilon_1 - i\epsilon_2 \qquad \ldots\ldots(3)$$

in such a way that the electric displacement

$$D = E + 4\pi P \qquad \ldots\ldots(4)$$

is given by the real part of $\epsilon E_0 e^{2\pi i\nu t}$. Then

$$4\pi P_1 = (\epsilon_1 - 1)E_0; \qquad 4\pi P_2 = \epsilon_2 E_0. \qquad \ldots\ldots(5)$$

If we assume that the magnetic field H does not act upon the electron, the energy absorbed by the adsorbed atoms per unit time and volume is, according to Maxwell's equations, given by

$$\bar{\dot{u}} = \frac{1}{4\pi}\overline{E\dot{D}}, \qquad \ldots\ldots(6)$$

where the bars indicate averages over one period.

If the number of atoms per cm² in the surface layer is denoted by z, and their number per unit volume by N, the rate of absorption per unit surface is given by $\dot{u}z/N$. Substituting into (6) from (2) and (4) and making use of (5) we find

$$\frac{z}{N}\bar{\dot{u}} = \frac{z}{N}\frac{\epsilon_2 \nu E_0^2}{4}. \qquad \ldots\ldots(7)$$

Thus the probability p for the absorption of light is given by

$$p = \frac{z}{N}\frac{\bar{\dot{u}}}{I} = \frac{z}{N}\frac{\epsilon_2 \nu E_0^2}{4I}, \qquad \ldots\ldots(8)$$

if I is the intensity of the incident light. Since the average values of electric and magnetic energies in a light wave are equal, we have

$$I = \frac{c}{8\pi}(\overline{E^2} + \overline{H^2}) = \frac{c\overline{E^2}}{4\pi} = \frac{cE_0^2}{8\pi}. \qquad \ldots\ldots(9)$$

Insertion of this into (8) leads to

$$p = \frac{z\,2\pi\nu\epsilon_2}{Nc}. \qquad \ldots\ldots(10)$$

Here it is assumed that the velocity of light in the surface layer is approximately equal to the velocity c in vacuum, and that the local electric field acting on an electron is equal to the field strength in vacuum. These assumptions are equivalent to the conditions

$$\epsilon_1 - 1 \ll 1, \qquad \epsilon_2 \ll 1, \qquad \ldots\ldots(11)$$

so that the interaction between the induced polarizations of different atoms can be neglected. It also excludes the possibility of appreciable reflexion from a supporting layer. Such terms would not, however, change the order of magnitude of (10).

The shape and magnitude of the absorption band is determined by ϵ_2. From very general arguments by S. Whitehead (1946) and B. Gross (1941) it follows, with the use of (11), that

$$\Delta\epsilon = \frac{2}{\pi}\int_0^\infty \epsilon_2(\nu)\,\frac{d\nu}{\nu}, \qquad \ldots\ldots(12)$$

where $\Delta\epsilon$ is the contribution of the oscillators to the static dielectric constant. If we assume that the absorption band is centred about the proper frequency ν_0 of the oscillators and is of a mean width $\Delta\nu$, the average value $\bar{\epsilon}_2$ is found from (12):

$$\Delta\epsilon \simeq \frac{2}{\pi}\,\frac{\Delta\nu}{\nu_0}\,\bar{\epsilon}_2, \qquad \ldots\ldots(13)$$

so that the average probability of absorption within the absorption range will be given, according to (10), by

$$\bar{p} = \frac{z}{N}\,\frac{\pi^2\nu_0{}^2\Delta\epsilon}{c\Delta\nu}. \qquad \ldots\ldots(14)$$

Now the restoring force of an elastically bound electron is $4\pi^2 m\nu_0{}^2\Delta x$, and thus the average displacement in a constant field E_0 is $\Delta x = eE_0/4\pi^2 m\nu_0{}^2$. Hence, since there are Nf such oscillators per unit volume,

$$\Delta\epsilon = \frac{4\pi e\Delta xNf}{E_0} = \frac{e^2Nf}{\pi\nu_0{}^2 m}. \qquad \ldots\ldots(15)$$

Inserting this into (14), we obtain

$$\bar{p} = \frac{\pi e^2 zf}{mc\Delta\nu} = \frac{\pi e^2\lambda_0 zf}{mc^2}\,\frac{\nu_0}{\Delta\nu}, \qquad \ldots\ldots(16)$$

where $\lambda_0 = c/\nu_0$.

Thus for an absorption band in the visible ($\lambda_0 \simeq 5\times 10^{-5}$ cm.) with a large oscillator strength ($f \simeq 1$) and a broadening $\Delta\nu/\nu_0 \simeq 1/2$, we obtain for a density of $z = 10^{15}$ atoms per unit surface ($e^2/m^2c = 3\times 10^{-13}$ cm.)

$$\bar{p} = 9\times 10^{-2},$$

i.e. absorption of about 10%.

§ 2. PHOTO-EMISSION

Emission of photoelectrons from adsorbed atoms can be obtained by the impact of light quanta with an energy less than the ionization energy of the free atom. The reason for this is that part of the ionization energy may be provided by the difference in the energies of adsorption of the neutral atom and the positive ion. Photo-emission may thus be considered as a two-stage process: (i) absorption of a quantum $h\nu$ in the absorption band centred around $h\nu_0$ leading to a transition of the electron from the ground state to an excited level; (ii) emission of the electron from the excited state.* Two mechanisms can be suggested for the second step. Either the energy difference Δ between the adsorption energies of the excited neutral atom and the positive ion is larger than the energy α required for the emission of an electron from the excited atom (and in this case electrons will be emitted with an average kinetic energy $\alpha - \Delta$), or else α is slightly smaller than Δ;

---

* A similar two-stage process has been suggested by Ryzhanov (1939); as the second step he assumes a transition of the excited metal electron into a lower state with a simultaneous excitation or emission of an electron in the supporting semi-conductor.

then the missing energy may be provided by the thermal energy of the supporting solid. This latter possibility is similar to the case of the internal photo-effect (cf. Mott and Gurney, 1940, p. 134) and, as in that case, a sudden drop in the photo-yield is to be expected as the temperature is lowered below a critical value. For this second mechanism the kinetic energy of the emitted electrons should be of the order of the thermal energy. For both mechanisms suggested, the selective photo-effect should be centred about a frequency corresponding to a strong absorption line of the free atom. Actually the maximum photo-yield of Cs atoms adsorbed on a layer of Cs_2O lies at 8000 A. (cf. Kluge, 1933; de Boer, 1935, p. 327), whereas the resonance line for the free Cs atoms form a doublet at about 8500 A. and 8900 A. For Rb on Rb_2O the corresponding figures are: 6500–6800 A. for the yield maximum and 7900 A. for the resonance line. This small shift of the maximum of the selective photo-effect can be expected in view of the different adsorption energies of the atoms in the ground state and excited state. For potassium the agreement is less satisfactory; the resonance line is at approximately 7700 A., whereas Kluge (1933) finds the maximum photo-yield at 4600–5200 A.; this peak, however, is not very pronounced, and its position cannot be accurately determined because of the proximity of a much stronger maximum at 4100 A., which is due to the supporting layer.

Another conclusion arising out of the present theory is that the spectral width of the emission band should increase proportional to $\sqrt{T}$, similar to the increase in the width of the absorption band in solids. Thus, the higher the temperature, the further should the photoelectric threshold be shifted towards longer wavelengths. No experimental evidence on this point is available at present.

REFERENCES

de Boer, J. H., 1935. *Electron Emission and Adsorption Phenomena* (Cambridge: The University Press).

Gross, B., 1941. *Phys. Rev.*, **59**, 748.

Kluge, W., 1933. *Phys. Z.*, **34**, 125.

Mott, N. F. and Gurney, R. W., 1940. *Electronic Processes in Ionic Crystals* (Oxford: The Clarendon Press).

Ryzhanov, S., 1939. *J. Exp. Theor. Phys., U.S.S.R.*, **9**, 38.

Whitehead, S., 1946. *Trans. Faraday Soc.* (in press).

CORRIGENDUM

"The Propagation of Supersonics in Capillary Tubes", by J. May (*Proc. Phys. Soc.*, **50**, 558 (1938)).

The figures in the seventh column of the table, headed "Amplitude absorption coefficient, Practical value", should be divided by 2.

(The author wishes to thank Mr. J. E. Drummond for pointing out an arithmetical error.)

FLUCTUATIONS IN STREAMS OF THERMAL RADIATION

By W. B. LEWIS, F.R.S.,
Telecommunications Research Establishment, Ministry of Supply

MS. received 13 *July* 1946

ABSTRACT. An ultimate limit to the sensitivity of a detector of thermal radiation is set by the inherent fluctuations in the stream of radiation. This paper presents the relevant quantitative relations to enable these fluctuations to be calculated in practical cases. The result is presented as the minimum detectable periodic modulation of the power of a stream for general conditions under which the detector may receive streams of radiation from surroundings at different temperatures and thermal equilibrium is not necessarily established. The modification necessary when the spectral frequency-distribution differs from that of full black-body radiation is also presented.

Limiting factors other than radiation fluctuations are not discussed.

§1. INTRODUCTION

CONSIDERATION of the fluctuations of radiation in a defined volume has played an important part in the history of the theory of radiation. It was shown by Einstein (1909) that the fluctuations could be represented as the sum of two terms, one which would be accounted for by classical wave-theory and the other having no interpretation on this theory but representing exactly the fluctuations expected on an extreme light-quantum view in which the quanta are assumed to behave as "classical" particles (e.g. Fowler, 1936, p. 765).

It is known, however, that an attempt to relate Planck's radiation law to such a view of a light-quantum or photon "gas" fails (e.g. Born, 1935, p. 215). The discrepancy has been traced to unwarranted assumptions such as the identity of individual quanta in the application of statistics, and when statistics omitting such assumptions are applied (Einstein–Bose statistics) to the photon gas, agreement with Planck's formula is achieved (see Born, *loc. cit.* p. 223). Moreover, deducing the fluctuations of radiation in a defined volume by applying Einstein–Bose statistics to a photon gas, the correct relation established by Einstein (1909) is obtained (Fürth, 1928; Fowler, *loc. cit.* p. 765).

It may therefore be assumed that full reliance may be placed on deductions of fluctuations in streams of radiation made by applying correct statistics to radiation treated as a photon gas.

It appears that although so much attention has been focused on radiation fluctuations, an explicit expression directly applicable to fluctuations limiting radiation pyrometry has not previously been published. It is the aim of this paper to set the theoretical knowledge of radiation fluctuations in the form most suited to the applied science of electronics by which such fluctuations may be observed.

In the field of electronics there is much experience of fluctuations or "noise" and the relevant theory is well developed; it will therefore be convenient to draw a parallel with the theory of the shot effect in presenting the results.

A radiation pyrometer may consist of a small thermal detecting element on to which the radiation is focused from a mirror. The equilibrium temperature of the detector will be determined by the balancing of the streams of radiation from the source via the mirror and from other surroundings against the heat conducted by the supports. It is assumed that the detector is vacuum mounted. Fluctuations of the detector temperature will arise from the fluctuations of the heat flow along these paths.

In critical comment it may be noted that as usual in practical considerations of fluctuations the assumption is being made that the whole apparatus may be divided into two parts. Discussion of fluctuations is related to one part while the other is assumed to obey the ordinary laws of large scale continuous physics (Fowler, *loc. cit.* p. 788). Here the division is being made at the boundary of the thermal detecting element. In practice it will almost inevitably be necessary to take account of fluctuations of voltage or current flow in the element, and also in the amplifier system used for detection. It is therefore desirable to present the results of this consideration of radiation fluctuations in such a form that these other fluctuations arising in the observing system may readily be incorporated.

To narrow the problem to determining the fluctuations of a stream of radiation, suppose the detector is idealized as an infinitesimal speck of matter enclosed in a small hollow sphere of volume v with diffusing perfectly reflecting internal walls in which there is a very small aperture of area a. The detector is thus an ideal black body of surface area a with a thermal capacity which is that of the enclosed radiation.

§ 2. FLUCTUATIONS IN A DEFINED VOLUME

The fluctuation of radiation energy in such an enclosure but without the speck of matter was the problem studied by Einstein in 1909. Writing η as the radiation energy of frequency ν to $\nu + d\nu$ in v, at any given instant η will differ from the mean η_0 by a small quantity ϵ so that $\eta = \eta_0 + \epsilon$.

It is necessary to assume that $v \gg \lambda^3$ so that the energy is statistically significant, where λ is the wave-length corresponding to ν.

Using the general statistical relations between entropy and probability, and determining the entropy of radiation over a frequency range ν to $\nu + d\nu$ from Planck's radiation formula, Einstein arrived at the result

$$\overline{\epsilon^2} = \eta_0 h\nu + \frac{c^3}{8\pi\nu^2 d\nu}\frac{\eta_0^2}{v}. \qquad \ldots\ldots(1)$$

Regarding the radiation as a light-quantum gas and writing n = number of quanta in volume v, we have

$$\eta_0 = nh\nu, \qquad \ldots\ldots(2)$$

so the first term of equation (1) is $n(h\nu)^2$. This is the fluctuation to be expected on the basis of classical statistics, but it is known that such statistics are not applicable to light-quanta in this way and Einstein–Bose statistics should be used instead.

From the theory of fluctuations in Einstein–Bose statistics, it is known that where in classical statistics the mean-square fluctuation of the number of particles $\Delta n^2 = n$, in the Einstein–Bose statistics the mean-square fluctuation

$$\Delta n^2 = n + \frac{n^2}{N}, \qquad \text{......(3)}$$

where N is the total number of independent cells which can be occupied by the particles, in this case the total number of independent standing wave vibrations in volume v.

This number is known from fundamental electro-magnetic wave theory (see e.g. Roberts, 1928, p. 398) to be

$$\frac{8\pi\nu^2 d\nu}{c^3} \cdot v,$$

so we have

$$\overline{\epsilon^2} = (h\nu)^2 \Delta n^2 = (\text{from eqn. (3)})\ n(h\nu)^2 + \frac{n^2(h\nu)^2 c^3}{8\pi\nu^2 d\nu} \cdot \frac{1}{v} \qquad \text{......(4)}$$

which by equation (2) may be seen to be identical with equation (1).

Noting that from Planck's radiation formula

$$\eta_0 = \frac{8\pi\nu^2 d\nu}{c^3} \cdot \frac{h\nu}{e^{h\nu/kT} - 1} \cdot v\ ; \qquad \text{......(5)}$$

the second term of equation (1) may be recast as

$$\frac{\eta_0 h\nu}{e^{h\nu/kT} - 1}\ ;$$

the relative magnitude of the two terms may be appreciated by rewriting equation (1) as

$$\overline{\epsilon^2} = \eta_0 h\nu \left[1 + \frac{1}{e^{h\nu/kT} - 1}\right]. \qquad \text{......(6)}$$

The second term is relatively large only when $h\nu \ll kT$, in which limit the expression tends to $\overline{\epsilon^2} = \eta_0 kT$.

§ 3. INTEGRATION OF FLUCTUATIONS OVER THE SPECTRUM

It is of interest to follow out the integration of $\overline{\epsilon^2}$ over all values of ν, because in practice it may be desired to appreciate the effect of absorption bands in parts of the spectrum.

By substituting the value of η_0 given by equation (5) in equation (6) we obtain

$$\overline{\epsilon^2} = \frac{8\pi\nu^2 d\nu}{c^3} \cdot \frac{(h\nu)^2 e^{h\nu/kT}}{(e^{h\nu/kT} - 1)^2} \cdot v. \qquad \text{......(7)}$$

Writing $\xi = h\nu/kT$, equation (7) becomes

$$\overline{\epsilon^2} = (kT)^5 \frac{8\pi v}{c^3 h^3} \cdot \frac{\xi^4 e^\xi}{(e^\xi - 1)^2} \cdot d\xi.$$

The total energy fluctuation is

$$\overline{\Delta E_v^2} = (kT)^5 \frac{8\pi v}{c^3 h^3} \int_0^\infty \frac{\xi^4 e^\xi}{(e^\xi - 1)^2} \cdot d\xi.$$

The function $\frac{\xi^4 e^\xi}{(e^\xi-1)^2}$ is shown in figure 1.

Write
$$y=\frac{\xi^4}{e^\xi-1};$$

then
$$\frac{dy}{d\xi}=\frac{4\xi^3}{e^\xi-1}-\frac{\xi^4 e^\xi}{(e^\xi-1)^2}.$$

$$\therefore\quad \overline{\Delta E_v^2}=(\boldsymbol{k}T)^5\frac{8\pi v}{\boldsymbol{c}^3\boldsymbol{h}^3}\left[\int_0^\infty -\frac{dy}{d\xi}\,.\,d\xi+\int_0^\infty\frac{4\xi^3}{e^\xi-1}\,.\,d\xi\right]$$

$$=(\boldsymbol{k}T)^5\frac{8\pi v}{\boldsymbol{c}^3\boldsymbol{h}^3}\left\{\left[\frac{-\xi^4}{e^\xi-1}\right]_0^\infty+\frac{4\pi^4}{15}\right\}.$$

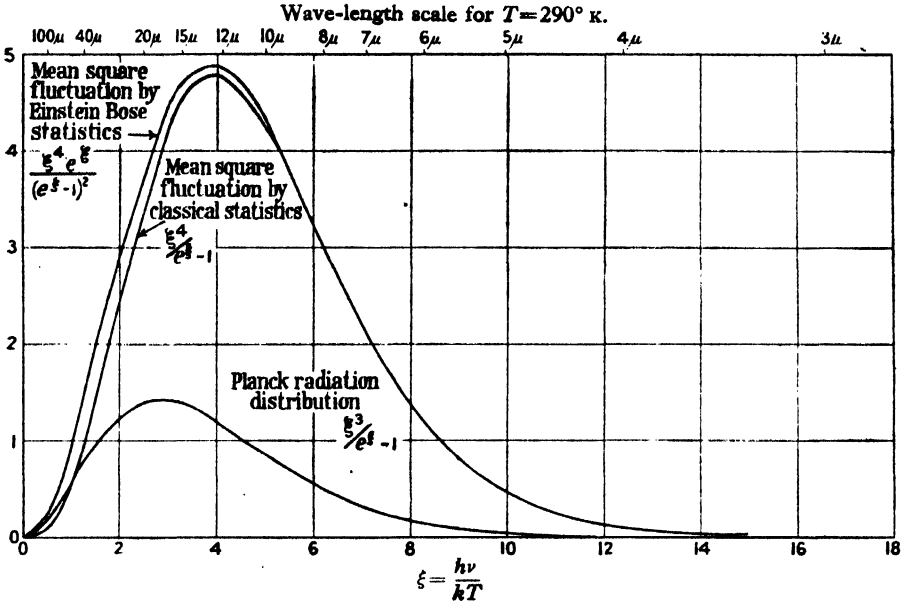


Figure 1.

The last term occurs in the integration of Planck's formula to relate it to Stefan's Law (see Wilson, 1944, p. 84), and since

$$\left[\frac{-\xi^4}{e^\xi-1}\right]_0^\infty=0,$$

we have

$$\overline{\Delta E_v^2}=\frac{(2\pi\boldsymbol{k}T)^5}{15\boldsymbol{c}^3\boldsymbol{h}^3}\,.\,v. \qquad \text{......(8)}$$

From standard works (e.g. Roberts, 1928, p. 369) we have $E_T=\frac{4\sigma T^4}{\boldsymbol{c}}$ where σ = Stefan's constant $=\frac{2\pi^5\boldsymbol{k}^4}{15\boldsymbol{c}^2\boldsymbol{h}^3}$ from the integration of Planck's formula, and E_T is the energy density of full black-body radiation of temperature T.

Equation (8) becomes, on substituting,

$$\overline{\Delta E_v^2}=4E_T\boldsymbol{k}Tv, \qquad \text{......(9)}$$

a relation which may be obtained directly from the general relation

$$\Delta E^2 = kT^2\frac{dE}{dT} \qquad \ldots\ldots(10)$$

(Fowler, *loc. cit.*, equation 2156, p. 764), noting that

$$\frac{dE_T}{dT} = \frac{16\sigma T^3}{c} = \frac{4E_T}{T}.$$

It may be of interest to note that if the second term of equation (1) were omitted we should have obtained instead of equation (9) $\overline{\Delta E_v^2} = 3{\cdot}83 E_T kTv$, and as already noted, the contribution of the second term is mainly in the low frequency or long wave-length region.

§4. FLUCTUATIONS OF RADIATION ENERGY CROSSING AN AREA

Regarding radiation as a light-quantum gas consisting of quanta of energy $h\nu$ moving in all directions with velocity c, then the number crossing an area a per second would be $\frac{1}{4}n_1 ca$ where n_1 is the number per unit volume. Applying Einstein-Bose statistics, the mean-square fluctuation $\overline{\Delta E_a^2}$ of the radiation of frequency ν crossing an area a would be

$$\tfrac{1}{4}n_1 ca(h\nu)^2\left[1+\frac{n_1}{N}\right].$$

Noting that $\eta_0 = n_1 h\nu v$, this may be written

$$\overline{\Delta E^2_{a\nu}} = \frac{ca}{4v}\eta_0 h\nu\left[1+\frac{1}{e^{h\nu/kT}-1}\right]$$

$$= \frac{ca}{4v}\overline{\epsilon^2} \text{ per second (from equation (6)).} \qquad \ldots\ldots(11)$$

The total energy fluctuation may be obtained by integrating (11) over all frequencies. By comparison with equation (8) this may be written down at once as

$$\Delta E_a{}^2 = \frac{ca}{4}\cdot\frac{(2\pi kT)^5}{15c^3h^3} \text{ or by equation (9)} = E_T kTca. \qquad \ldots\ldots(12)$$

§5. TIME VARIATION OF FLUCTUATIONS IN A RADIATION STREAM

Equation (12) is a relation closely parallel with the basic idea of the shot-effect fluctuations of the saturation current in a thermionic diode. If λ is the mean rate of arrival of electrons at the anode, the mean-square fluctuation of the charge reaching the anode per second is $\overline{q^2} = e^2\lambda$ where e is the electronic charge. But $e\lambda = I_0$, the mean current, so that

$$\overline{q^2} = eI_0. \qquad \ldots\ldots(13)$$

In practical considerations of shot noise in circuits it is convenient to use a related equation

$$\overline{i_f^2} = 2eI_0\Delta f \qquad \ldots\ldots(14)$$

where $\overline{i_f^2}$ is the mean square current fluctuation over the frequency range f to

$f\Delta+f$. The transition from equation (13) to equation (14) has been established by a necessarily elaborate mathematical argument, and it may be taken over for the thermal case. When applied to equation (12) we obtain

$$\overline{\Delta W_{af}^2}=2\boldsymbol{k}TE_T\boldsymbol{c}a\Delta f \qquad \ldots\ldots(15)$$

where W_a is the rate of flow of energy or power crossing a.

It is to be understood that this may be applied to determine the frequency spectrum of the fluctuations generated in a body exposed to this stream of radiation.

Moreover, if the body is black and at temperature T and therefore radiating an equivalent stream to balance the received radiation, then the effective mean square fluctuation of energy flow from the body over the frequency range Δf is doubled, viz.,

$$\overline{\Delta W_{bf}^2}=4\boldsymbol{k}TE_T\boldsymbol{c}a\Delta f, \qquad \ldots\ldots(16)$$

or, since $E_T\boldsymbol{c}=4\sigma T^4$, where $\sigma=$ Stefan's constant,

$$\overline{\Delta W_{bf}^2}=16\sigma\boldsymbol{k}T^5a\Delta f$$

or

$$\sqrt{\overline{\Delta W_{bf}^2}}=4T^2\sqrt{a\boldsymbol{k}T\Delta f}, \qquad \ldots\ldots(17)$$

a relation established by Daunt (1945)* for the minimum power detectable as a sinusoidal variation of frequency f by an isolated body connected to its surroundings only by radiation.

§6. RADIATION FLUCTUATIONS IN PRACTICE

The above argument by which this relation has been reached indicates how this minimum detectable power may be determined when the detector receives streams of energy from surroundings at differing temperatures and thermal equilibrium is not necessarily assumed. Using equation (15) for each stream, the total mean-square fluctuation may be determined by summing these contributions for all the streams.

It may be noted that in terms of σ equation (15) becomes

$$\overline{\Delta W_{af}^2}=8\sigma\boldsymbol{k}T^5a\Delta f. \qquad \ldots\ldots(18)$$

An object at temperature T_1 subtending a small solid angle α in a direction making an angle θ with the normal to the plane of a will contribute a term

$$\overline{\Delta W_{af_1}^2}=32\sigma\boldsymbol{k}T_1{}^5a\cos\theta\Delta f\,.\,\frac{\alpha}{4\pi}. \qquad \ldots\ldots(19)$$

The magnitude of this contribution may be assessed, taking $\sigma=5{\cdot}65\times10^{-5}$ erg. cm.$^{-2}$ sec.$^{-1}$ deg.$^{-4}$, $\boldsymbol{k}=1{\cdot}38\times10^{-16}$ erg. deg.$^{-1}$, $a=1$ mm., $\Delta f=1$ c./s., and writing $T_0=290°$ K., we have

$$\overline{\Delta W_{af_1}^2}=0{\cdot}97\times10^{-24}\,\alpha\cos\theta(T_1/T_0)^5 \text{ watt}^2.$$

The minimum detectable power as defined by Daunt (1945) will be

$$\sqrt{\sum_{n=1}^{n=n}\overline{W_{af_n}^2}}=0{\cdot}98\times10^{-12}\left\{\pi\left(\frac{T}{T_0}\right)^5+\sum_{n=1}^{n=n}\alpha_n\cos\theta_n\left(\frac{T_n}{T_0}\right)^5\right\}^{\frac{1}{2}} \text{ watts}$$

where the summation is carried out over all objects in view, and T is the temperature of the detector. This assumes that the detector and all objects are perfectly

---

* Also independently by Milatz (1943).

black. Modifications for grey and reflecting objects will be obvious, but it may be noted that if the object reflects radiation back to the detector, neither the outgoing nor the reflected stream contributes to the fluctuations if the distance of the reflector is $\ll c/\Delta f$.

It may be noted for comparison that the total radiation crossing a, namely $\sigma a T_0^4 = 4 \times 10^{-4}$ watts, while for $T = T_0$ from equation (17) the minimum detectable power for $\Delta f = 1$ c./s. is $2{\cdot}46 \times 10^{-12}$ watts. On the view presented in this paper, this minimum detectable power is determined by the statistical variation in the number of quanta crossing a per second. The average energy of the quanta is about $4kT$ and the number n_2 crossing a per second is about 10^{16}. The fluctuation of the number is about $\sqrt{n_2} \approx 10^8$, so the minimum detectable power is about 10^{-8} of the total flow of energy in this example.

If the radiation in any stream is not full but the spectral distribution is known to be $f(\nu)E$ (where E represents the spectral distribution for full black-body radiation) the resultant fluctuations may be calculated by the following relation (from equations (7), (11), (15) and (19)):

$$\overline{\Delta W^2_{af_1}} = \frac{4(\boldsymbol{k}T_1)^5}{\mathbf{c}^2\boldsymbol{h}^3} \, . \, a \cos\theta \, . \, \Delta f \, . \, \alpha \int_0^\infty \frac{(\boldsymbol{h}\nu/\boldsymbol{k}T_1)^4 e^{h\nu/kT_1}}{(e^{h\nu/kT_1} - 1)^2} \, . \, f(\nu) \, \frac{\boldsymbol{h}\,d\nu}{\boldsymbol{k}T_1} \qquad \ldots\ldots(20)$$

In this expression the integral is a numerical term which, if $f(\nu) = 1$, we have seen is $4\pi^4/15$; it must not, however, be assumed that its value is independent of T_1 (for in practice $f(\nu)$ cannot in general be expressed as a function of ν/T_1).

ACKNOWLEDGMENT

I would like to thank Dr. R. A. Smith, Dr. J. G. Daunt and Dr. R. Kompfner for much helpful discussion in the preparation of this paper.

REFERENCES

Born, M., 1944. *Atomic Physics* (London: Blackie and Son).
Daunt, J. G., 1945. *S.R.E. Report*, C. L. Misc., **52**. (Unpublished.)
Daunt, J. G., 1947. (In course of publication).
Einstein, A., 1909. *Phys. Z.*, **10**, 185.
Fowler, R. H., 1936. *Statistical Mechanics* (Cambridge: The University Press).
Fürth, R., 1928. *Z. Phys.*, **50**, 310.
Lorentz, H. A., 1916. *Les Théories Statistiques en Thérmodynamique* (Leipzig: Teubner).
Milatz, J. H. W., 1943. Brownsche Beweging. *Proceedings of the XXIXth Netherland Scientific Congress.*
Roberts, J. K., 1928 *Heat and Thermodynamics* (London: Blackie and Son.)
Wilson, H. A., 1944. *Modern Physics* (London: Blackie and Son).

REFLECTING MICROSCOPES

By C. R. BURCH, F.R.S.,
University of Bristol

MS. received 5 *March* 1946 ; *in revised form* 16 *September* 1946

ABSTRACT. The history of reflecting microscopes is reviewed, and the leading Schwarzschild design formulae quoted. Two reflecting objectives are described, and photomicrographs taken with one of these are reproduced.

§ 1. HISTORICAL INTRODUCTION

REFLECTING microscopes are almost as old as reflecting telescopes, for it was early realized that if the light be sent backwards through a telescope objective, it becomes a microscope objective, albeit of inconveniently great focal length. Newton made a reflecting microscope objective consisting of an ellipsoidal mirror and a diagonal flat; a two-mirror objective more akin to the Cassegrain type is described in Smith's *Complete System of Optics* (1738). These "compound reflecting engiscopes"—as such microscopes were called—were not aplanatic, so that although they could give good images at the low numerical aperture of 0·05 – 0·1 (which for many years satisfied astronomers in their corresponding telescope objectives) they gave seriously comatic off-axis images when the N.A. was raised. Further, a large fraction of the N.A. was necessarily obstructed by the shadow of the second mirror : this could in certain circumstances lead to undesirable effects such as a spurious doubling of the number of lines in the image of a grating. Finally, no satisfactory technique had been developed, either for making or testing the required aspheric surfaces. Accordingly, reflecting micro-objectives passed into oblivion. However, in 1905, Schwarzschild gave the analytical solution of the problem of designing an aplanatic (i.e. spherically corrected and coma-free) two-mirror telescope objective. Algebraically this is of course also the solution of the problem of the two-mirror micro-objective when used at infinite or very great tube length, though the range of numerical values of Schwarzschild's two design parameters m, ϵ, corresponding to practical designs of micro-objective, differs from that corresponding to practical designs of telescope objective. Schwarzschild's classic memoir seems to have attracted relatively little attention, for in 1922 the problem of the two-mirror aplanat was attacked independently by Chrétien, at the instigation of G. W. Ritchey, who had been struck by the fact that the 60″ telescope at Mount Wilson was more nearly aplanatic when used as a Cassegrain than as a Newtonian. "He suspected," writes Chrétien, "that the introduction of the hyperbolic mirror produced a kind of compensation of the aberration of the parabolic mirror : he asked me to study this system theoretically, and in particular, to enquire if it was not possible to improve the imaging properties further by freely forsaking the parabolic and hyperbolic shapes previously given to mirrors."

One of the first to take up afresh the study of reflecting microscopes seems to have been D. D. Maksutov, who, in U.S.S.R. patent No. 40859 (1932), mentions the use of the sphere-cardioid aplanatic pair as a micro-objective,[1] and shows an extremely ingenious "solid" reflecting objective in which the same air-glass surface, after acting as "first mirror" by total internal reflection, lets out the image formed by the second (silvered) surface at normal incidence. His work does not appear to have been suggested by that of Schwarzschild or Chrétien, whose general formulae are not quoted.

More recently he has designed reflecting objectives composed of two spherical mirrors with one or more relatively weak lenses on the image side. Photomicrographs taken in U.V. with one of these objectives have been published by E. M. Brumberg and others. The combination of reflection and refraction has been exploited in a different way by B. K. Johnson, who has published U.V. photomicrographs taken with a reflecting objective analogous to the Mangin mirror, while E. H. Linfoot has combined more reflection with less refraction in two reflecting objectives of Schmidt type, which he showed at the Physical Society's exhibition in 1939.

§ 2. SCHWARZSCHILD APLANATS

Consider the reflecting objective shown in figure 1, consisting of a concave mirror of paraxial radius ρ and a convex of radius R separated by a distance ϵ, the unit of length being the focal length of the combination. Parallel light is imaged at distance m from the pole of the ρ-mirror. ρ, R, ϵ, m are all positive in the diagram. Schwarzschild showed that the shapes of the two mirrors may be so defined that the system is spherically corrected and satisfies the sine condition, and he solved the resulting differential equations in finite form, obtaining for the ρ-mirror the equation

$$\frac{1}{r}=\frac{t}{\epsilon}+\frac{1}{m}\,\frac{\left(1-\dfrac{t}{\epsilon}\right)^{\frac{1}{1-\epsilon}}}{(1-t)^{\frac{\epsilon}{1-\epsilon}}} \qquad \ldots\ldots(1)$$

where

$$t=\sin^2 \tfrac{1}{2}u \qquad \ldots\ldots(2)$$

and r is the length of a ray leaving the object-point at angle u to its incidence-point on the ρ-mirror.

The incidence point of this ray on the second mirror may conveniently be expressed in Cartesian coordinates:

$$\left.\begin{aligned} x&=\epsilon-(1-t)\left[\frac{r(\epsilon-2t)+t}{\epsilon-rt}\right]\ldots\\ y&=2t^{\frac{1}{2}}(1-t)^{\frac{1}{2}}=\sin u\ldots \end{aligned}\right\} \qquad \ldots\ldots(3)$$

The incidence angle i' of this ray on the first (ρ) mirror is given by

$$\tan i'=\frac{(\epsilon-r+1-t)}{\epsilon-t}\left(\frac{t}{1-t}\right)^{\frac{1}{2}}. \qquad \ldots\ldots(4)$$

Equations (1) to (4) define the aplanat in terms of m and ϵ, so that we have to choose numerical values for these parameters.

I do not propose to discuss the design-problem exhaustively, but shall give, for the convenience of those wishing to work on these objectives, the principal approximate formulae with which the designer will find himself concerned. We may be influenced by the coefficient of astigmatism or that of field curvature or by the obstruction ratio or by manufacturing considerations. Schwarzschild showed that the astigmatic interfocal distance which would be needed in object-space to give parallel rays at (small) angle θ to the axis in image-space is

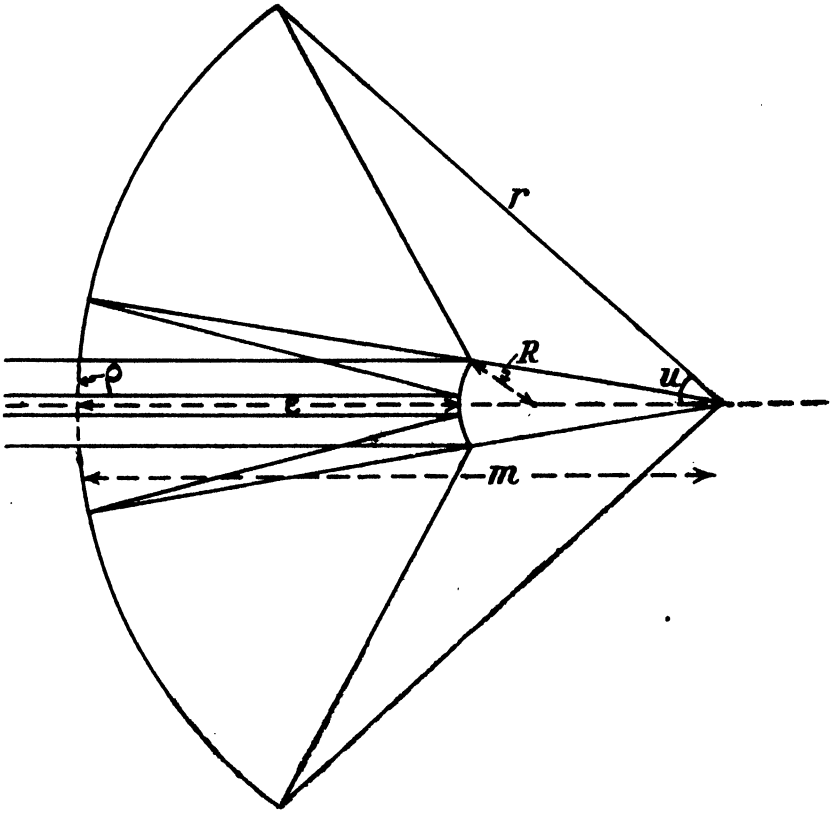


Figure 1. Schwarzschild aplanat.

$$\text{A.I.D.} = \frac{2-\epsilon}{2m} \cdot \theta^2, \qquad \ldots\ldots(5)$$

that is, $\frac{2-\epsilon}{2m}$ times the "thin lens" value. He showed also that the surface midway between the focal lines will be curved

$$\frac{(m-1)^2+\epsilon(1-\epsilon)}{2m\epsilon} \cdot \theta^2 \qquad \ldots\ldots(6)$$

away from the Gaussian image plane.

The fraction of the N.A. shadowed out by the R-mirror is approximately

$$\frac{1}{m-\epsilon} \quad \text{or} \quad m-\epsilon, \qquad \ldots\ldots(7)$$

whichever is less than unity. (In the latter case the light must first pass through a hole in the R-mirror). This approximation contains no unused margin allowance, and is a little over optimistic when the N.A. is not small.

For ease of manufacture we would prefer one mirror to be spherical if this is permissible. Accordingly we seek series approximations to equations (1) and (3). Schwarzschild gives for the ρ-mirror the Cartesian series

$$x = m + \left[\frac{1-m}{\epsilon} - 1\right]\frac{y^2}{4m} - \left\{\frac{(1-m)^2}{\epsilon} - \frac{(1-m)}{\epsilon} + \frac{1}{2\epsilon}\right\}\frac{y^4}{16m^3}$$
$$+\left[2\left(\frac{1-m}{\epsilon}\right)^3 - 2\left(\frac{1-m}{\epsilon}\right)^2 + \frac{2}{\epsilon}\left(\frac{1-m}{\epsilon}\right) - \frac{1+\epsilon}{6\epsilon^2}\right]\frac{y^6}{64m^5}\text{, etc.} \qquad \ldots\ldots(8)$$

If we disregard terms above y^6, the "non-elliptic" part of this is

$$P_{NE}(x) = \frac{-m}{128\epsilon^2}\left[\frac{1+\epsilon}{3} + \frac{\epsilon}{1-m-\epsilon}\right]\left(\frac{y}{m}\right)^6, \qquad \ldots\ldots(9)$$

while the remainder is an ellipse of paraxial radius

$$\rho = \frac{2m\epsilon}{m+\epsilon-1} \qquad \ldots\ldots(10)$$

and of eccentricity e' given by

$$e'^2 = \left(\frac{1-m+\epsilon}{1-m-\epsilon}\right)^2 + \frac{2\epsilon^2}{(1-m-\epsilon)^3}, \quad = e_0'^2 - \frac{1}{4\epsilon}\left(\frac{\rho}{m}\right)^3, \qquad \ldots\ldots(11)$$

where e_0' is the eccentricity of that ellipse whose foci are the object-point and paraxial intermediate image-point.

For the R-mirror Schwarzschild gives $x = A + By^2 + Cy^4 + Dy^6 + Ey^8$ where

$$\left.\begin{aligned} &A = m-\epsilon\,;\quad B = \frac{1-m}{4\epsilon}\,;\quad C = -\frac{1}{8}\cdot\frac{m}{4\epsilon}\,;\\ &D = -\frac{1}{96}\cdot\frac{1+4\epsilon}{\epsilon}\cdot\frac{m}{4\epsilon}\,;\quad E = -\frac{1}{1536}\cdot\frac{2+11\epsilon+30\epsilon^2}{\epsilon^2}\cdot\frac{m}{4\epsilon}.\end{aligned}\right\} \qquad \ldots\ldots(12)$$

This is approximately an ellipse of paraxial radius

$$R = \frac{2\epsilon}{m-1} \qquad \ldots\ldots(13)$$

and of eccentricity e given by

$$e^2 = 1 + \frac{2\epsilon^2 m}{(1-m)^3}. \qquad \ldots\ldots(14)$$

The non-elliptic part of the sixth power coefficient is

$$P_{NE}(D) = D - \frac{2C^2}{B}. \qquad \ldots\ldots(15)$$

Let us summarize the main design-restrictions which these formulae imply. For greatest ease of manufacture we equate e' and e simultaneously to zero. This gives $\epsilon = 2$, so that the system is anastigmatic, and $m = 2 + \sqrt{5}$, $\rho = 1 + \sqrt{5}$; $R = \sqrt{5} - 1$, a monocentric objective. The price that we must pay for having zero first coefficient of asphericity on both mirrors—so that they may be spherical up to N.A. approaching 0·5—is the high intrinsic obstruction ratio, $1/\sqrt{5}$. If we wish to retain anastigmatism, we must set $\epsilon=2$, and, except in the case just considered, e' and e both differ from zero for every value of m, so that both mirrors must be aspherized. If we are prepared to give up exact anastigmatism, and be content with aplanatism only, we may set e zero, giving $\epsilon^2=\frac{(m-1)^3}{2m}$, and select m so as to give as small an obstruction ratio, $\frac{1}{m-\epsilon}$, as we please. This gives an aplanat in which—provided terms involving the sixth power of the N.A. may be neglected—the R-mirror may be spherical. Alternatively we can set e' zero, but this leads to a higher astigmatic coefficient for a given obstruction ratio. I show in a separate paper that when the design is optimized, the R-mirror may be spherical in an objective of small obstruction ratio, up to 0·65 N.A. If we ask—may the R-mirror ever be exactly spherical, up to N.A. unity, the answer is yes, provided $\epsilon=\frac{1}{2}$, $m=2$. The Schwarzschild concave curve then becomes a cardioid, and the combination is the well-known sphere-cardioid pair used in aplanatic dark-ground condensers, usually in approximation as two spheres.

§ 3. EXPERIMENTAL

I have made two "through-type" reflecting objectives, which, though they are corrected for finite tube-lengths, may be regarded as approximating to Schwarzschild aplanats having respectively $\rho=4{\cdot}31$ cm., $R=0{\cdot}85$ cm., $m=8$, $\epsilon=4$, $f=0{\cdot}75$ cm.; N.A. ·58, tube-length 32 cm., magnification ×47, and $\rho=5$ cm., $R=0{\cdot}397$ cm., $m=20$, $\epsilon=13$, $f=0{\cdot}3$ cm.; N.A. ·65, tube-length 30 cm., magnification × 100. To this objective I have added a normal-incidence oil-immersion lens, the surface of which is spherical and concentric with the axial object point. This raises the N.A. to 0·98, and the magnification to about 152, the magnification now being proportional to the refractive index of the lens.

The first objective has both mirrors aspherized: the concave mirror was first aspherized so as to annul spherical aberration for a certain arbitrarily chosen separation between the mirrors and object position, after which the convex mirror was moved axially with respect to the concave so as to annul off-axis coma (the final image position being unchanged). The spherical aberration re-introduced by this change was then annulled by aspherizing the convex mirror. This procedure produces a sufficient approximation to aplanatism: a more elaborate procedure would be needed for N.A.$>$0·7.

Local figuring was needed to restore revolution symmetry lost during aspherizing, and the residual error is of "cobbled pavement" type, of the order of ⅛ fringe by double transmission. The leading off-axis error for image points 1 cm. off-axis is a fraction of a fringe of astigmatism.

The second objective has a nominally spherical convex mirror, the spacing between the mirrors being adjusted to give the best compromise between high-order and Seidel coma; the leading error for image points ½ cm. off-axis consists of a fraction of a fringe of compromise—comatic retardation of the edge of the wave-front remote from the axis: for image points 1 cm. off-axis the leading error is astigmatism. This objective also has a small fraction of a fringe of "cobbled pavement" error.

I find it advisable, with both objectives, to stop out, in the substage condenser, that part of the N.A. which is obstructed in the objective. If this is not done, the contrast is poor when the substage N.A. is reduced, as is to be expected, for if the substage N.A. is reduced until only the cone obstructed by the objective is supplied, the conditions are those of "patch-stop dark-ground illumination" and black objects are seen bright on a black ground. [If a dark-ground effect is desired, I find it better to use a standard dark-ground illuminator supplying a cone lying outside that used by the objective, as the diffraction rings surrounding images are then much less prominent.]

The visual performance of both objectives used with full substage aperture on stained specimens seems comparable in contrast and resolving power with that of refracting objectives of equal N.A.

Some idea of the photographic performance may be obtained from figures 2–6, which are reproduced from enlargements made by Mr. H. Busby of films taken by Dr. G. P. Occhialini with the second objective. For their kindly interest and enthusiastic co-operation I record my grateful thanks.

Figures 2 and 3, taken without the oil immersion component, show "thorium stars" × 1140, and treponema pallida, silver-stained by Levaditi's method, × 950, 0·65 N.A. ; 0·65 substage N.A.

Figures 4, 5 and 6 were taken with oil immersion : figure 4 shows part of the long chromosome of Drosophila, × 1420, ·98 N.A., ·45 substage N.A. ; figure 5 shows treponema pallida, Levaditi stained, × 2230, and figure 6 gonococci in pus cells, stained with methylene blue, × 2230, 0·98 N.A., 0·98 substage N.A. The light source in all cases was a half-watt projector lamp, diffused by ground glass, with a green filter. No eyepiece was used, the film being placed at the primary image.

REFERENCES

BRUMBERG, 1939. *C.R. Acad. Sci., U.S.S.R.*, **25**, 473 ; 1941 a. *Ibid.*, **31**, 658 ; 1941 b. *Ibid.*, **32**, 486 ; 1943. *Nature, Lond.*, **152**, 357.
BURCH, 1943. *Nature, Lond.*, **152**, 748.
BURCH, 1943. *Proc. Phys. Soc.*, **55**, 433. 1945. *Ibid.*, **57**, 567.
CHRÉTIEN, 1922. *Rev. d'Optique*, **1** 232.
CLAY, 1939. *J. Sci. Instrum.*, **16**, 49.
JOHNSON, 1941. *Proc. Phys. Soc.*, **53**, 714.
LINFOOT, 1938. *J. Sci. Instrum.*, **15**, 405.
MAKSUTOV. U.S.S.R. Patent No. 40859.
SCHWARZSCHILD, 1905. *Theorie der Spiegelteleskop* (Göttingen Observatory).
SMITH, 1738. *Complete System of Optics.*

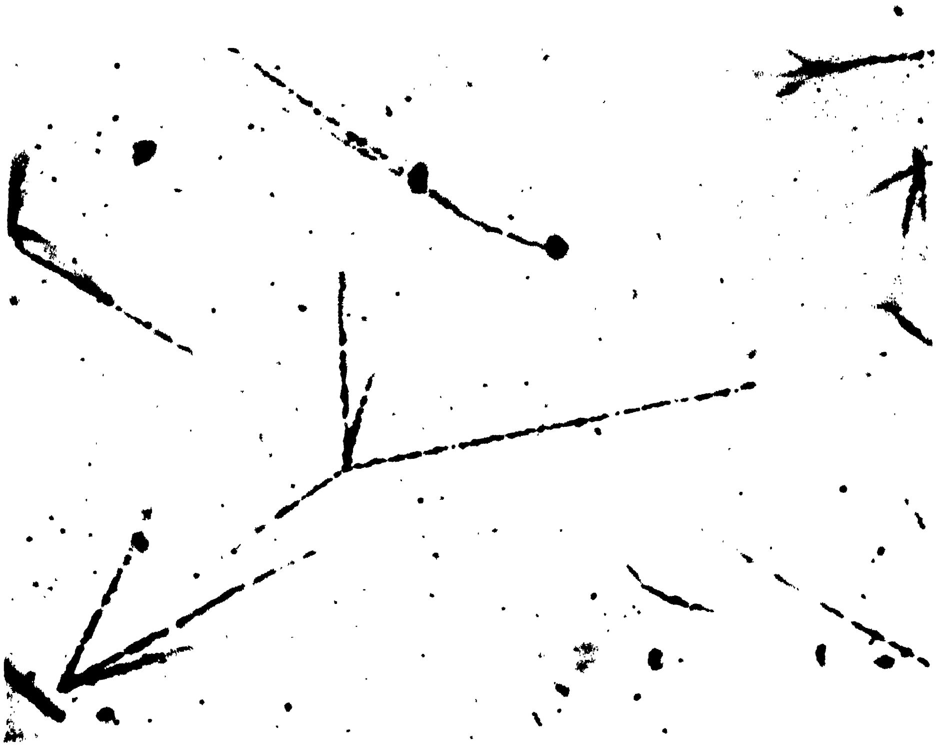


Figure 2.

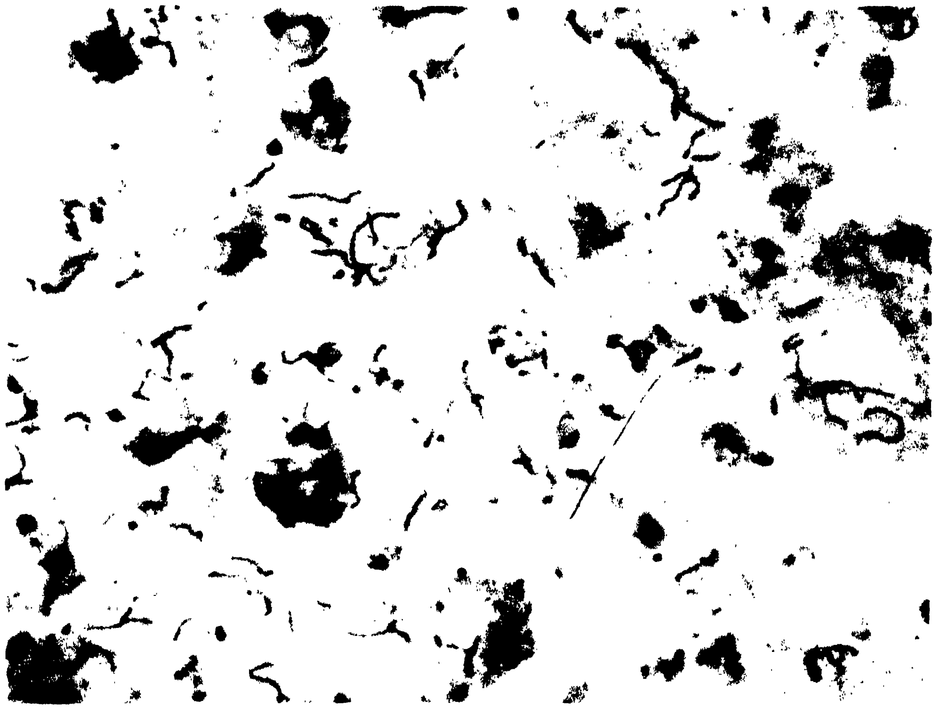


Figure 3.

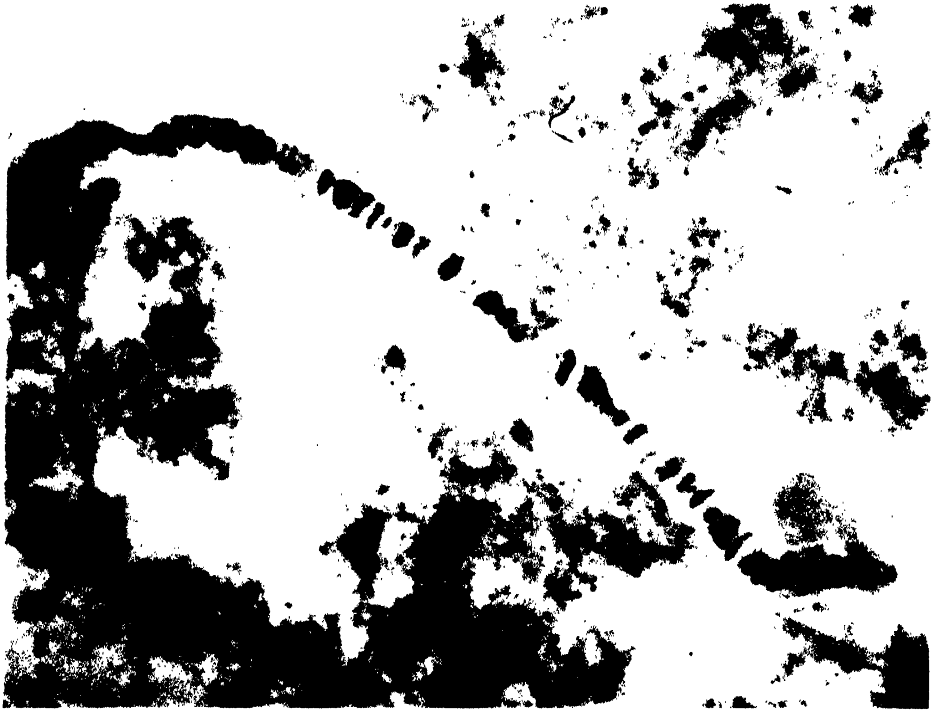


Figure 4.

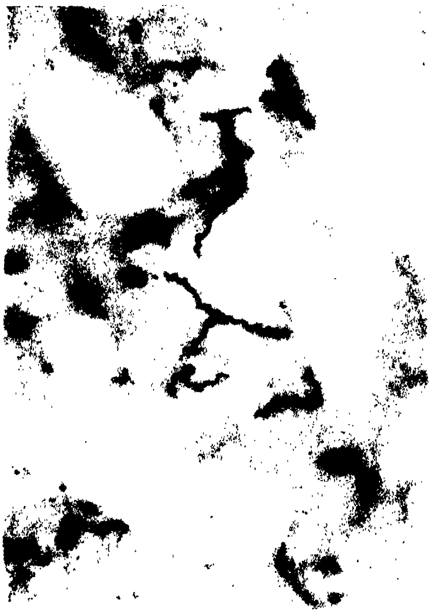


Figure 5.

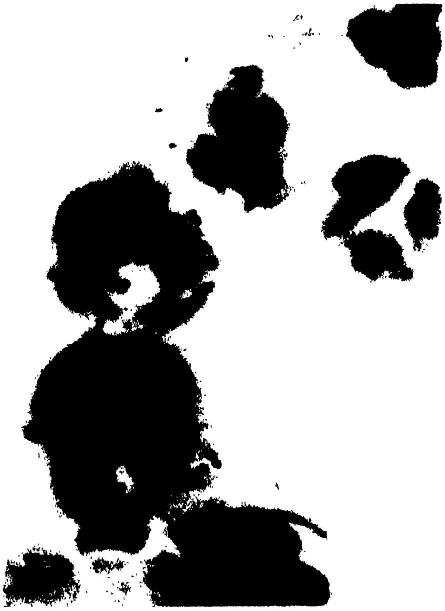


Figure 6.

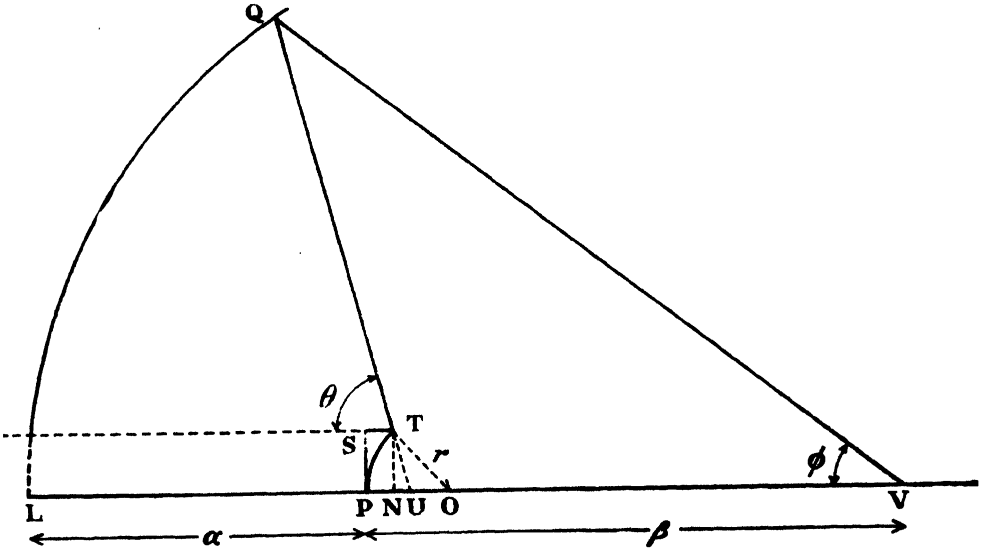
SEMI-APLANAT REFLECTING MICROSCOPES

By C. R. BURCH, F.R.S.,
Bristol University

MS. received 5 *March* 1946 : *in revised form* 9 *August* 1947

ABSTRACT. Numerical values are derived for the offence against the sine condition of reflecting micro-objectives composed of one aspheric concave mirror and one spherical convex mirror. Numerical apertures up to 0·65 are found practicable.

The purpose of this paper is to derive numerical values for the offence against the sine condition ("OSC") of "semi-aplanat" reflecting microscope objectives composed of an aspheric concave "primary" mirror and a spherical convex "secondary", placed so as to form the final image at infinite or very great tube length.



Semi-aplanat reflecting micro-objective.

Such an objective is shown in the figure. We suppose that the concave mirror is figured so that the final image is free from spherical aberration. Then the path length VQTS, from the object-point V to a point S on a plane through the pole of the convex mirror, is independent of the position of the incidence-point, Q, on the concave mirror. Further, if r be the radius of the convex mirror, θ the angle of the intermediate ray, $LP=\alpha$, $PV=\beta$, e the eccentricity and a the semi-major axis of the ellipse of which Q is a point and U and V foci, we have (see figure)

$$VQTS=\beta+2\alpha, \qquad \ldots\ldots(1)$$

$$\frac{r}{2}\left(2-\sec\frac{\theta}{2}\right)=\mathrm{PU}, \qquad \ldots\ldots(2)$$

$$r\cos\frac{\theta}{2}=\mathrm{NO}, \qquad \ldots\ldots(3)$$

$$\beta-\frac{r}{2}\left[2-\sec\frac{\theta}{2}\right]=\mathrm{VU}=ae, \qquad \ldots\ldots(4)$$

and

$$\beta+2\alpha-\frac{r}{2}\left[2-2\cos\frac{\theta}{2}-\sec\frac{\theta}{2}\right]=\mathrm{VL}+\mathrm{LP}-\mathrm{PU}+\mathrm{NO},=\beta+2\alpha-\mathrm{ST}+\mathrm{TU}$$
$$=\mathrm{VQ}+\mathrm{QU}=2a. \qquad \ldots\ldots(5)$$

Hence

$$e=\frac{\beta-\frac{r}{2}\left[2-\sec\frac{\theta}{2}\right]}{\beta+2\alpha-\frac{r}{2}\left[2-2\cos\frac{\theta}{2}-\sec\frac{\theta}{2}\right]}. \qquad \ldots\ldots(6)$$

From the geometry of the ellipse,

$$\sin\phi=\frac{2(1-e^2)\cos\frac{\theta}{2}\sin\frac{\theta}{2}}{(1+e)^2-4e\sin^2\frac{\theta}{2}}. \qquad \ldots\ldots(7)$$

The height of the parallel ray imaged at angle ϕ is

$$r\sin\frac{\theta}{2}, \qquad \ldots\ldots(8)$$

Set

$$K(\theta)=\frac{(1-e^2)\cos\frac{\theta}{2}}{(1+e)^2-4e\sin^2\frac{\theta}{2}}. \qquad \ldots\ldots(9)$$

Then the offence against the sine condition is

$$\mathrm{OSC}=\frac{\sin\phi_{\text{ideal}}-\sin\phi_{\text{actual}}}{\sin\phi_{\text{ideal}}}=\frac{K(0)-K(\theta)}{K(0)}. \qquad \ldots\ldots(10)$$

Substituting for e from (6) we can expand $K(\theta)$ as a power series in θ^2 with leading terms

$$K(\theta)=\frac{1-e_0}{1+e_0}\left[1+\frac{\theta^2}{8}\left\{\frac{e^2_0-6e_0+1}{(1+e_0)^2}+\frac{re_0}{(1-e_0)\left(\beta-\frac{r}{2}\right)}\right\}\right] \ldots\ldots(11)$$

so that if the θ^2 coefficient of $K(\theta)$ is to be zero—i.e. if the objective is to have zero Seidel coefficient of coma—we must have

$$\frac{\beta}{r}=\frac{1}{2}+\frac{e_0(1+e_0)^2}{(1-e_0)(-e_0^2+6e_0-1)}. \qquad \ldots\ldots(12)$$

We would like to have $\beta \gg r$, so as to have a negligible fraction of the N.A. obstructed by the shadow of the convex mirror. Consider therefore first the limiting case $\beta/r=\infty$, which implies $e_\theta=e_0,=(1+e_0^2)/6,=3-2\sqrt{2}=\cdot 1718\ldots$. The magnification of the intermediate image, $(1-e_0)/(1+e_0)$, then becomes $1/\sqrt{2}$.

$K(\theta)$ now becomes

$$\frac{\cos\frac{\theta}{2}}{\sqrt{2}\left(1-\frac{1}{2}\sin^2\frac{\theta}{2}\right)}=\frac{1}{\sqrt{2}}\left(1-\frac{\theta^4}{128}+\text{etc.}\right) \qquad \ldots\ldots(13)$$

so that if we set the OSC tolerance at 1/400 and neglect higher terms, we must limit θ to roughly $1/\sqrt{2}$ radian, and the NA ($\sin\phi$) to about 1/2. We can reduce the maximum OSC by a re-choice of e_0, provided we do not demand that the θ^2-coefficient of OSC be zero. When $\beta/r=\infty$, we have for any e_0

$$K(\theta)=\frac{1-e_0}{1+e_0}\cdot\frac{\cos\frac{\theta}{2}}{1-\frac{4e_0}{(1+e_0)^2}\sin^2\frac{\theta}{2}} \qquad \ldots\ldots(14)$$

so that we should choose e_0 so as to make $1-\frac{4e_0}{(1+e)^2}\sin^2\frac{\theta}{2}$ imitate $\cos\frac{\theta}{2}$ as closely as possible. Let us make the imitation exact at $\theta=60°$; then we must set $e_0=\cdot1896\ldots$, and the imitation is found to be correct to about 1·1 parts in 400 up to $\theta=60°$, for which $\sin\phi=\cdot6812$, which is also the value of $(1-e_0)/(1+e_0)$, the magnification of the intermediate image. This suggested that a useful approximation to optimal design for N.A. 0·65 or so might be obtained when β/r is finite by a corresponding increase in e_0—say, a 10% increase—over the value given by (12). Three values of e_0 were accordingly chosen in this way for the three values 5·75, 3·50 and 2·75 of β/r and the OSC was computed for a number of values of θ by means of (6), (9), and (10). The results, together with the obstructed fraction of the N.A., i.e. $\frac{(1+e_0)}{2(1-e_0)}\cdot\frac{r}{\beta}$ and the associated values of α/r, are given in table 1.

Table 1

θ	$\sin\phi$	OSC	$\sin\phi$	OSC	$\sin\phi$	OSC	$\sin\phi$	OSC
10	0·1188	$0{\cdot}27\times10^{-3}$	0·1163	$0{\cdot}23\times10^{-3}$	0·1138	$0{\cdot}22\times10^{-3}$	0·1114	$0{\cdot}21\times10^{-3}$
20	0·2368	0·99	0·2317	0·88	0·2269	0·82	0·2222	0·81
30	0·3533	1·90	0·3457	1·68	0·3385	1·61	0·3314	1·63
40	0·4617	2·55	0·4571	2·40	0·4476	2·29	0·4383	2·39
50	0·5771	2·26	0·5647	2·19	0·5531	2·33	0·5417	2·70
55	0·6300	1·46	0·6166	1·62	0·5904	1·92	0 5917	2·50
60	0·6812	0 0	0·6670	0·44	0·6540	0·66	0·6405	1·94
65	0·7304	−2·21	0·7154	−1·37	0·7014	0·34	0·6875	0·91
70	0·7771	−5·47	0·7617	−4·04	0·7471	−2·71	0·7328	−0·68
e_0	0·1896		0·20		0·21		0·2201	
β/r	inf		5·75		3·50		2·75	
α/r	inf		10·0		5·143		3·486	
Obstruction	0·0		$\frac{1}{7{\cdot}66}$		$\frac{1}{4{\cdot}57}$		$\frac{1}{3{\cdot}5}$	

These objectives should therefore be reasonably coma-free up to 0·65 N.A. It is fortunate that the sign of the OSC is such that it can be corrected by a suitable turn-*down* of the edge of the convex mirror, together with a corresponding reduction of asphericity of the concave.

THE HEATING OR COOLING OF A SOLID SPHERE IN A WELL-STIRRED FLUID

By S. PATERSON,
Imperial Chemical Industries Ltd., Explosives Division

MS. received 26 *August* 1946

ABSTRACT. A solid sphere of uniform initial temperature is heated or cooled in a finite mass of fluid, well stirred and externally insulated. Temperature continuity is assumed at the surface. Solutions are presented suitable for all values of time, radius and conductivity, and of the ratio of heat capacities of sphere and fluid. Numerical results are given, and certain mathematical relations noted.

---

§ 1. SERIES SOLUTION

A SOLID sphere of radius a, diffusivity κ, heat capacity c_1 and temperature zero is plunged at time $t=0$ into a mass of fluid of heat capacity c_2 at temperature T_0. The surface temperature of the sphere is assumed equal at all subsequent times to that of the fluid, which is so well stirred or of such high conductivity that its temperature is always uniform throughout. The outer boundary of the fluid is impervious to heat.

Then the temperature T at a distance R from the centre of the sphere is determined by

$$\left.\begin{aligned} &\frac{\partial}{\partial t}(RT)=\kappa\frac{\partial^2}{\partial R^2}(RT), \quad R<a, \quad t>0,\\ &T=0, \quad R<a, \quad t=0,\\ &T=T', \quad R=a, \quad t>0,\\ &T'=T_0, \quad t=0,\\ &\frac{3\kappa c_1}{a}\frac{\partial T}{\partial R}=-c_2\frac{dT'}{dt}, \quad R=a, \quad t>0, \end{aligned}\right\} \quad \ldots\ldots(1)$$

where $T'(t)$ is the temperature of the fluid.

For convenience, let $R/a\equiv r$, $\kappa t/a^2\equiv \tau$, $RT/aT_0\equiv u$, $T'/T_0\equiv u'$; also $c_1/c_2\equiv w$. Then the equations are

$$\left.\begin{aligned} &\frac{\partial u}{\partial \tau}=\frac{\partial^2 u}{\partial r^2}, \quad r<1, \quad \tau>0, \qquad u=u', \quad r=1, \quad \tau>0,\\ &\qquad\qquad\qquad\qquad\qquad\qquad u'=1, \quad \tau=0,\\ &u=0, \quad r<1, \quad \tau=0, \qquad u=0, \quad r=0, \quad \text{all } \tau,\\ &3w\left(\frac{\partial u}{\partial r}-\frac{u}{r}\right)+\frac{du'}{d\tau}=0, \quad r=1, \quad \tau>0. \end{aligned}\right\} \quad \ldots\ldots(2)$$

If $p\equiv -s^2$ is the Heaviside operator $\partial/\partial\tau$, these equations transform into

$$\left.\begin{aligned} &\frac{\partial^2 u}{\partial r^2}=-s^2u, \qquad u=0, \quad r=0,\\ &3w\left(\frac{\partial u}{\partial r}-\frac{u}{r}\right)-s^2(u-1)=0, \quad r=1. \end{aligned}\right\} \quad \ldots\ldots(3)$$

The solution is readily found to be

$$u=\frac{s^2\sin rs}{3w(\sin s-s\cos s)+s^2\sin s} \qquad \ldots\ldots(4)$$

$$\equiv f(s)/F(s), \text{ say.}$$

The poles s_n of u are the roots of

$$s\cot s=1+s^2/3w, \qquad \ldots\ldots(5)$$

and apart from a simple pole at zero consist of real and distinct pairs, positive and negative. The solution (4) therefore transforms into

$$u=\mathcal{L}_{s\to 0}\frac{f(s)}{F(s)}+2\sum_{n=1}^{\infty}\frac{f(s_n)}{s_nF'(s_n)}e^{-s_n^2\tau}.$$

Hence $$u=\frac{r}{w+1}+2\sum_1^{\infty}\frac{\sin rs_n}{\sin s_n}\cdot\frac{e^{-s_n^2\tau}}{3(w+1)+s_n^2/3w}, \qquad \ldots\ldots(6)$$

where s_n is the nth positive non-zero root of (5). The corresponding fluid temperature is

$$u'=(u)_{r=1}=\frac{1}{w+1}+2\sum_1^{\infty}\frac{e^{-s_n^2\tau}}{3(w+1)+s_n^2/3w}. \qquad \ldots\ldots(7)$$

The above formal solution can be justified by verifying that (6) and (7) do in fact satisfy (2).

When $\tau=\infty$, $u=u'r=r/(w+1)$, as is required. The fraction F of the final heat transfer which has taken place by time t is therefore

$$F=\frac{1-u'}{1-1/(w+1)}, \qquad \ldots\ldots(8)$$

that is.

$$F=1-\frac{2}{3w}\sum_1^{\infty}\frac{e^{-s_n^2\tau}}{1+s_n^2/9w(1+w)}, \qquad \ldots\ldots(9)$$

(5), (6) and (7) correspond, for the present condition of uniform initial temperature, to the general series solution given by Peddie (1901). When the fluid has infinite heat capacity, that is, when the surface temperature of the sphere is maintained at T_0, they reduce to

$$u=r+\frac{2}{\pi}\sum_1^{\infty}(-1)^n\frac{\sin n\pi r}{n}e^{-n^2\pi^2\tau}, \qquad \ldots\ldots(10)$$

$$u'=1,$$

which is a familiar case. (9) then becomes

$$F=1-\frac{6}{\pi^2}\sum_1^{\infty}\frac{e^{-n^2\pi^2\tau}}{n^2}. \qquad \ldots\ldots(11)$$

The nth root s_n of (5) can be seen from a graph of the functions $\tan s$ and $s/(1+s^2/3w)$ to lie between $n\pi$ and $(n+\frac{1}{2})\pi$. For $n<(\sqrt{3w}/\pi)-\frac{1}{2}$, s_n approaches $(n+\frac{1}{2})\pi$ as n is increased, thereafter returning towards $n\pi$; thus, if w is small, s_n approaches $n\pi$ from the outset. s_n can be calculated without difficulty by successive approximation. If n/w is sufficiently large, the following expansion is also useful:

$$s_n = n\pi + \frac{3w}{n\pi} - \frac{9w^2(w+2)}{(n\pi)^3} + \dots \qquad \dots\dots(12)$$

The first three roots, which are more than we shall require, are presented in table 1 for a wide range of w.

Table 1. The first three positive roots of $s \cot s = 1 + s^2/3\omega$

w	s_1	s_2	s_3	w	s_1	s_2	s_3
∞	4·712	7·854	10·996	5·0	4·236	7·296	10·329
1000	4·492	7·722	10·899	2·0	3·972	6·938	9·940
500	4·490	7·720	10·895	1·0	3·726	6·681	9·714
400	4·487	7·719	10·893	0·5	3·506	6·502	9·576
300	4·486	7·716	10·890	0·2	3·312	6·376	9·487
200	4·485	7·713	10·885	0·1	3·233	6·330	9·454
100	4·479	7·702	10·866	0·05	3·188	6·307	9·439
50	4·464	7·674	10·830	0·02	3·161	6·293	9·430
20	4·421	7·601	10·759	0·01	3·152	6·290	9·426
10	4·352	7·490	10·573	0·0	3·142	6·283	9·425

The series in (6), (7), etc. no doubt always converge. The rapidity of convergence, however, varies markedly with τ. Thus in (7), if $\tau \geqslant 1$, the exponent $-s_n^2\tau$ increases numerically by a factor of at least $(2n+1)\pi^2$, and the entire series is negligible, so that $u = r/(w+1)$, $u' = 1/(w+1)$ and $F = 1$, irrespective of w; but when $\tau < 1$, the speed of convergence falls off very quickly, at least two terms being required for $\tau = 0{\cdot}1$, eleven for $\tau = 0{\cdot}01$ and over one hundred for $\tau = 0{\cdot}001$. Since for certain purposes it may be necessary to consider values of τ as low as, say, 10^{-10}, the series solution evidently cannot be regarded in practice as covering more than a fraction of the range.

§2. ALTERNATIVE SOLUTIONS

We return to (4), substitute $-q^2$ for s^2, and expand in powers of e^{-q}. This gives

$$u = \frac{q^2}{q^2 + 3wq - 3w}\{e^{-(1-r)q} - e^{-(1+r)q} + Qe^{-(3-r)q} - Qe^{-(3+r)q} + \dots\} \qquad \dots\dots(13)$$

and

$$u' = \frac{q^2}{q^2 + 3wq - 3w}\{1 + (Q-1)e^{-2q} + Q(Q-1)e^{-4q} + \dots\}, \qquad \dots\dots(14)$$

where

$$Q \equiv \frac{q^2 - 3wq - 3w}{q^2 + 3wq - 3w}. \qquad \dots\dots(15)$$

Since e^{-2q} transforms into $1 - \operatorname{erf} 1/\sqrt{\tau}$, we may expect that the bracketed terms beyond the first two in (13) and the first in (14) will be negligible when $1 - \operatorname{erf} 1/\sqrt{\tau}$ is small, say $<0{\cdot}1\%$ for $\tau \leqslant 0{\cdot}1$. Then (13) and (14) should be suitable for calculation precisely in the range in which (6) and (7) are unsuitable.

In order to obtain an approximation for sufficiently small τ, we may expand $q^2/(q^2+3wq-3w)$ in descending powers of q, and interpret operators of the type $q^{-n}e^{-Aq}$ in terms of Hartree's repeated error function integrals (Hartree, 1936). Thus, if $P\equiv 3w$,

$$\frac{q^2 e^{-(1\pm r)q}}{q^2+Pq-P} = \Phi_0\left(\frac{1\pm r}{2\sqrt{\tau}}\right) - P(2\sqrt{\tau})\Phi_1\left(\frac{1\pm r}{2\sqrt{\tau}}\right) + (P+P^2)(2\sqrt{\tau})^2\Phi_2\left(\frac{1\pm r}{2\sqrt{\tau}}\right) - \dots, \qquad \dots\dots(16)$$

where

$$\Phi_n(x) \equiv \int_x^\infty \Phi_{n-1}(\xi)\,d\xi$$

and

$$\Phi_0(x) \equiv 1 - \operatorname{erf} x.$$

When $r=1$, the terms in $1+r$ are relatively small, and the first approximation to u' is

$$u' = 1 - P.2\sqrt{\tau/\pi} + (P+P^2)\tau - \dots \qquad \dots\dots(17)$$

For sufficiently small $w\sqrt{\tau}$, say $<10^{-2}$, these expansions are useful, particularly (16), when $r \ll 1$, but they do not cover the required range. We shall therefore derive alternative expressions.

The terms in $e^{-(1\pm r)q}$ from (13) transform into

$$\mp \frac{1}{\pi i}\int_M \frac{e^{\mu^2\tau-(1\pm r)\mu}\mu\,d\mu}{(\mu-\alpha)(\mu-\beta)},$$

respectively, where α, β are the roots of $\mu^2+3w\mu-3w=0$, and M is a path in the μ-plane from $\infty e^{-i\pi/4}$ to $\infty e^{i\pi/4}$ passing a finite distance to the right of $|\alpha|$ and $|\beta|$. The terms therefore yield

$$u \simeq \frac{1}{\alpha-\beta}\{\alpha[E_{1-r}(\alpha)-E_{1+r}(\alpha)] - \beta[E_{1-r}(\beta)-E_{1+r}(\beta)]\}, \qquad \dots\dots(18)$$

where

$$\left.\begin{aligned} E_y(x) &\equiv \frac{1}{\pi i}\int_{M'} \frac{e^{\mu^2\tau - y\mu}\,d\mu}{\mu - x} \\ &= e^{x^2\tau - xy}\left[1 - \operatorname{erf}\left(\frac{y}{2\sqrt{\tau}} - x\sqrt{\tau}\right)\right] \\ &= e^{-y^2/4\tau}G\left(\frac{y}{2\sqrt{\tau}} - x\sqrt{\tau}\right), \end{aligned}\right\} \qquad \dots\dots(19)$$

where

$$G(z) \equiv e^{z^2}(1 - \operatorname{erf} z).$$

(18) and (19) give a first approximation to u. It can be further simplified for particular ranges of r or w. Thus, if r is small, (18) becomes

$$u \simeq \frac{2r}{\alpha-\beta}\left\{\alpha^2 E_1(\alpha) - \beta^2 E_1(\beta) + \frac{\alpha-\beta}{\sqrt{\pi\tau}}e^{-1/4\tau}\right\}. \qquad \dots\dots(20)$$

The relative temperature u/r at the centre is obtained at once from (20). Again, if r is nearly 1, the first bracketed term in (13) will, for small τ, be large compared with the second; so that

$$u \simeq \frac{1}{\alpha-\beta}\{\alpha E_{1-r}(\alpha)-\beta E_{1-r}(\beta)\}. \qquad \ldots\ldots(21)$$

A first approximation to the fluid temperature is obtained from the first term of (13), or by setting $r=1$ in (21). Thus

$$u' \simeq \frac{1}{\alpha-\beta}\{\alpha E_0(\alpha)-\beta E_0(\beta)\}. \qquad \ldots\ldots(22)$$

Finally, when $w=0$,

$$\begin{aligned} u &\simeq E_{1-r}(0)-E_{1+r}(0) \\ &= \operatorname{erf}\frac{1+r}{2\sqrt{\tau}}-\operatorname{erf}\frac{1-r}{2\sqrt{\tau}}, \qquad \ldots\ldots(23) \\ F &= \frac{3(q\cosh q-\sinh q)}{q^2\sinh q} \\ &= -3\tau+6\sqrt{\tau/\pi}+12\sqrt{\tau}\sum_1^\infty \Phi_1\left(\frac{n}{\sqrt{\tau}}\right). \qquad \ldots\ldots(24) \end{aligned}$$

(23) and (24) correspond to (10) and (11). (24) converges very rapidly indeed for $\tau \leqslant 0{\cdot}1$, giving $F=6\sqrt{\tau/\pi}-3\tau$ to better than 1 in 30,000, while for $\tau \geqslant 0{\cdot}1$ two terms at most are required of the series in (11).

No matter how small $1-r$ is, short of being absolutely zero, the expressions given by (18), (21) and (23) tend to zero with τ, as required. On the other hand, (22), which is of course continuous with (21) for $\tau>0$, tends to unity as τ approaches zero. The solutions thus correctly reproduce the required continuity for $\tau>0$ and discontinuity at $\tau=0$. It is not the least advantage of the operational method, as remarked by Jaeger (1945), that this is made possible in cases like the present without any special analytical device.

It is of interest to confirm that the second term in (14) is in fact negligible. This term transforms into

$$-\frac{6w}{\pi i}\int_M \frac{e^{\mu^2\tau-2\mu}\mu^2 d\mu}{(\mu-\alpha)^2(\mu-\beta)^2}.$$

We may resolve the integrand into partial fractions, and integrate the terms in $(\mu-\alpha)^{-2}$ and $(\mu-\beta)^{-2}$ by parts. There results

$$+\frac{12w}{(\alpha-\beta)^3}\Big\{[\alpha\beta-\alpha^2(\alpha-\beta)(\alpha\tau-1)]E_2(\alpha)-[\alpha\beta+\beta^2(\alpha-\beta)(\beta\tau-1)]E_2(\beta) \\ -(\alpha-\beta)(\alpha^2+\beta^2)\sqrt{\frac{\tau}{\pi}}e^{-\frac{1}{\tau}}\Big\}. \qquad \ldots\ldots(25)$$

Let $\alpha \equiv \tfrac{1}{2}(-3w+\sqrt{9w^2+12w})$ and $\beta \equiv \tfrac{1}{2}(-3w-\sqrt{9w^2+12w})$. Clearly $0<\alpha<1$ and $\beta<0$. Values of α and β for a wide range of w are shown in table 2. If, then, $\tau<0{\cdot}1$, $1/\sqrt{\tau}-\alpha\sqrt{\tau}$ and $1/\sqrt{\tau}-\beta\sqrt{\tau}$ are of the order of 3, or greater. Hence, using the asymptotic series for erf, we have

$$\left.\begin{aligned} E_2(\alpha) &\sim e^{-1/\tau}/\sqrt{\pi}(1/\sqrt{\tau}-\alpha\sqrt{\tau}), \\ E_2(\beta) &\sim e^{-1/\tau}/\sqrt{\pi}(1/\sqrt{\tau}-\beta\sqrt{\tau}). \end{aligned}\right\} \qquad \ldots\ldots(26)$$

Table 2. The roots α, β of $x^2+3wx-3w=0$

w	α	β	w	α	β
1000	0·9997	−3001	5·00	0·9410	−15·94
500	0·9993	−1501	2·00	0·8730	− 6·873
400	0·9992	−1201	1·00	0·7913	− 3·791
300	0·9989	− 901·0	0·50	0·6861	− 2·186
200	0·9983	− 601·0	0·20	0·5307	− 1·131
100	0·9967	− 301·0	0·10	0·4179	− 0·7179
50	0·9934	− 151·0	0·05	0·3195	− 0·4695
20	0·9839	− 60·98	0·02	0·2168	− 0·2768
10	0·9693	− 30·97	0·01	0·1583	− 0·1883

The factor $e^{-1/\tau}$ already guarantees that the term in $E_2(\alpha)$ in (23) is negligible; and the same can at once be shown to be true of the remaining two terms in virtue of the cancellation of the potentially large β^3.

Thus, for $\tau \leqslant 0{\cdot}1$, u' is given to a close approximation by (22), that is by

$$u' \simeq \frac{1}{\alpha-\beta}\{\alpha e^{\alpha^2\tau}(1+\operatorname{erf}\alpha\sqrt{\tau})-\beta e^{\beta^2\tau}(1+\operatorname{erf}\beta\sqrt{\tau})\}, \qquad \ldots\ldots(27)$$

and F follows from (8).

By an argument precisely similar to the above, it can be shown that for $\tau \leqslant 0{\cdot}1$ the terms in (13) beyond the first two are also negligible. So also, of course, is the second itself if r is near to 1.

§ 3. NUMERICAL RESULTS

u' and F have been computed for $w=1000$ to $w=0$ and $\tau=10^{-12}$ to $\tau=1$. (27) was used for $\tau \leqslant 0{\cdot}1$, and (7) for $\tau \geqslant 0{\cdot}1$. A comparison between the two types of solution is thus provided at $\tau=0{\cdot}1$. It will be seen from table 3 that the agreement is close, considering that only four-figure tables were used. The course of F is shown in figure 1. Here an additional check was made at $w=0$, $\tau=0{\cdot}1$ and 0·01, (11) yielding $F=0{\cdot}7705$, 0·3084 and (24) $F=0{\cdot}7705$, 0·3085. Finally, as an illustration, the temperature field was calculated for $w=1$, $\tau=10^{-4}$, 10^{-2}, 10^{-1} by (18) and for $\tau=0{\cdot}1$, 1 by (6). Very close agreement was found at $\tau=0{\cdot}1$. The distribution is shown in figure 2. Except at the centre u/r rises temporarily above its final value.

§ 4. APPENDIX

Several interesting relations emerge from the above. Thus, from (9), since $F \to 0$ as $\tau \to 0$,

$$\sum_{n=1}^{\infty}\frac{1}{s_n^2+9w(1+w)}=\frac{1}{6(1+w)}, \qquad \ldots\ldots(28)$$

s_n being the nth positive non-zero root of (5). $\Sigma\, 1/n^2=\pi^2/6$ is of course a special case, given by (11).

Again, from (11) and (24),

$$\sum_1^{\infty}\frac{e^{-n^2\pi^2\tau}}{n^2}=\frac{\pi^2}{6}\left\{1+3\tau-6\sqrt{\frac{\tau}{\pi}}-12\sqrt{\tau}\sum_1^{\infty}\Phi_1\left(\frac{n}{\sqrt{\tau}}\right)\right\}. \qquad \ldots\ldots(29)$$

Table 3. $u' \equiv T'/T_0$ for various values of $w \equiv c_1/c_2$ and $\tau \equiv \kappa t/a^2$

	Calculated from Equation (27)												Calculated from Equation (7)	
w \ τ	10^{-12}	10^{-11}	10^{-10}	10^{-9}	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-1}	$\geqslant 1$
1000·00	0·997	0·989	0·967	0·901	0·735	0·442	0·178	0·060	0·019	0·006	0·002	0·001	0·001	0·000
500·00		0·995	0·983	0·949	0·851	0·629	0·322	0·117	0·038	0·012	0·005	0·002	0·002	0·002
400·00		0·996	0·987	0·959	0·878	0·683	0·370	0·145	0·048	0·016	0·006	0·003	0·003	0·002
300·00			0·990	0·969	0·906	0·746	0·457	0·188	0·063	0·021	0·008	0·004	0·004	0·003
200·00			0·993	0·979	0·936	0·818	0·568	0·272	0·093	0·031	0·011	0·005	0·005	0·005
100·00			0·997	0·988	0·967	0·901	0·735	0·443	0·182	0·062	0·022	0·011	0·011	0·010
50·00				0·995	0·984	0·949	0·852	0·630	0·325	0·122	0·044	0·021	0·021	0·020
20·00					0·993	0·979	0·936	0·818	0·571	0·276	0·108	0·052	0·052	0·048
10·00					0·996	0·990	0·967	0·902	0·737	0·451	0·203	0·100	0·100	0·091
5·00						0·995	0·983	0·949	0·852	0·638	0·351	0·185	0·185	0·167
2·00							0·994	0·980	0·937	0·822	0·596	0·369	0·370	0·333
1·00							0·997	0·989	0·967	0·904	0·755	0·548	0·548	0·500
0·50								0·995	0·983	0·950	0·864	0·714	0·714	0·667
0·20									0·994	0·980	0·941	0·865	0·864	0·833
0·10									0·997	0·990	0·970	0·928	0·928	0·909
0·05										0·995	0·985	0·963	0·963	0·952
0·02											0·994	0·985	0·985	0·980
0·01											0·997	0·992	0·992	0·990

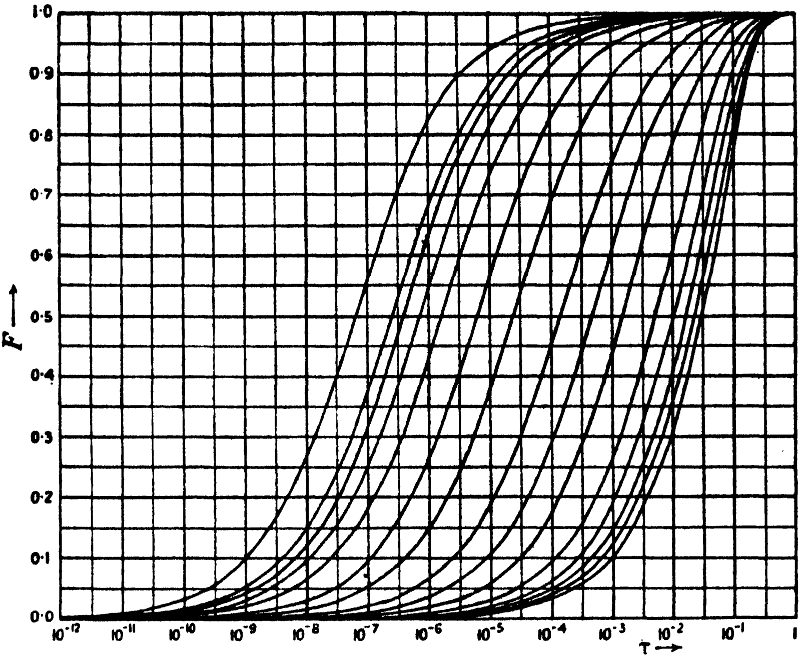


Figure 1. Variation of F with τ for various values of w. From left to right the curves correspond to $w = 1000$, 500, 400, 300, 200, 100, 50, 20, 10, 5, 2, 1, 0·5, 0·2, 0.

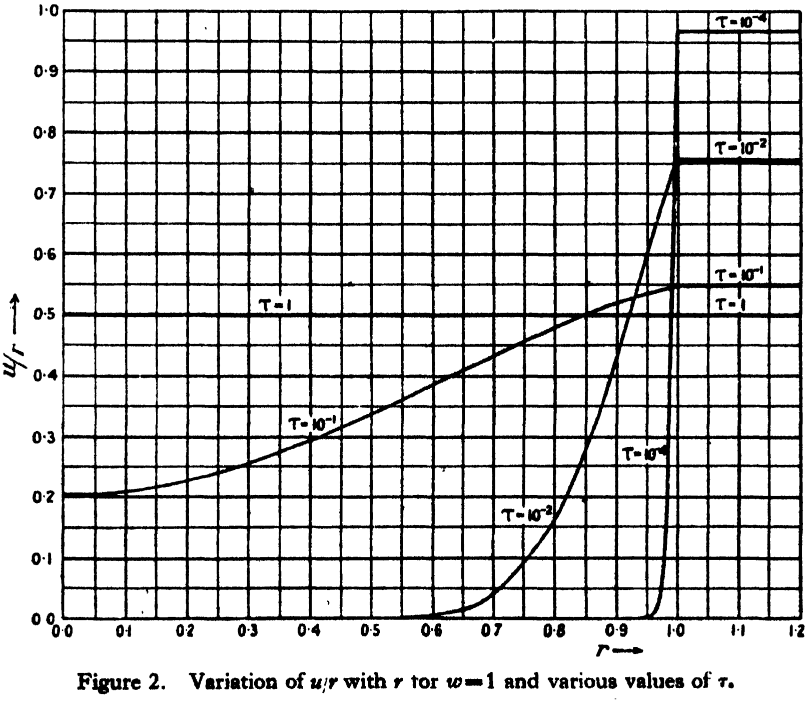


Figure 2. Variation of u/r with r for $w = 1$ and various values of τ.

This yields, very closely, for $0<x\leqslant 1$,

$$\sum_1^\infty \frac{e^{\;n^2x}}{n^2} \simeq \frac{\pi^2}{6} - \sqrt{\pi x} + \frac{x}{2} \qquad \ldots\ldots(30)$$

and

$$\sum_1^\infty \Phi_1(nx) \simeq \frac{1}{4x} - \frac{1}{2\sqrt{\pi}} + \frac{x}{12}. \qquad \ldots\ldots(31)$$

By differentiating (29) and (30) with regard to x, or integrating from x to ∞, a series of useful formulae can be derived. For example,

$$\sum_1^\infty e^{-n^2x} = \tfrac{1}{2}\left(\sqrt{\frac{\pi}{x}} - 1\right) + \sqrt{\frac{\pi}{x}} \sum_1^\infty e^{-n^2\pi^2/x} \qquad \ldots\ldots(32)$$

and

$$\sum_1^\infty \frac{e^{-n^2x}}{n^4} = \frac{\pi^4}{90} - \frac{\pi^2 x}{6} + \frac{2\sqrt{\pi}}{3}x^{3/2} - \frac{x^2}{4} + \pi(2\sqrt{x})^3 \sum_1^\infty \Phi_3\left(\frac{n\pi}{\sqrt{x}}\right), \qquad \ldots\ldots(33)$$

the series terms on the right of (32) and (33) being negligible for $0<x\leqslant 1$. (32) is a particular case of " Poisson's Identity ".

REFERENCES

Hartree, D. R., 1936. *Mem. Manch. Lit. Phil. Soc.*, **80**, 85.
Jaeger, J. C., 1945. *Proc. Camb. Phil. Soc.*, **41**, 43.
Peddie, W., 1901. *Proc. Edin. Math. Soc.*, **19**, 34.

THE APPLICATION OF IONOSPHERIC DATA TO RADIO COMMUNICATION PROBLEMS: PART II

By Sir EDWARD APPLETON, F.R.S. and W. J. G. BEYNON*
* Radio Division, National Physical Laboratory; now at
University College, Swansea

MS. received 30 *July* 1946 .† *Read* 15 *November* 1946

ABSTRACT. Graphs are given from which may be estimated the maximum usable frequency of radio waves reflected by an ionospheric layer in oblique incidence transmission. The curves based on the theory given in Part I of the paper are drawn for such ranges of layer thickness and layer height as are met with in practice. The limitations in the accuracy and applicability of the theory in practice are briefly discussed. Attention is also drawn to the occurrence of abnormal transmission conditions under which long-distance communication via the ionosphere is possible on frequencies exceeding the normally predicted values.

§ 1. INTRODUCTION

In Part I of this paper (Appleton and Beynon, 1940) we described a method of estimating the frequency range of radio waves which are deviated back to the ground when incident obliquely on the ionosphere, and which are, therefore, suitable for communication over various distances. For this purpose formulae

---

† Sections 1 to 7 of this paper were originally prepared in June 1942 and circulated at that time as a confidential paper by the Radio Research Board.—Authors.

were derived giving directly the value of the maximum usable frequency which is reflected by a thick ionized layer surrounding a curved earth, the distribution of electronic density in the layer being taken as "parabolic". In this continuation of the paper we exhibit the somewhat complicated formulae previously obtained for the maximum usable frequency in a graphical form, embodying a range of the relevant parameters which our practical experience suggests as adequate. Also, since the solution previously obtained involves certain approximations, we supplement its graphical expression by a discussion of the accuracy which can be attained in typical practical cases. Attention is also drawn to the manner in which the analysis can be expressed in the form of transmission curves suitable for the direct estimation of the maximum usable frequency from the vertical-incidence relation between equivalent height of reflection and frequency. In the final section of the paper we discuss the practical significance of abnormal *E*-layer reflections in effecting long-distance transmission on frequencies exceeding the normally predicted values.

§2. MAXIMUM USABLE FREQUENCY FACTORS

In the theoretical treatment of Part I of this paper the characteristics of the deviating layer in the ionosphere were specified in terms of three quantities, viz.:

(*a*) f^0, the ordinary ray critical frequency of the ionized layer for waves incident normally on it;

(*b*) h_0, the height above ground of the lower edge of the layer; and

(*c*) y_m, the vertical semi-thickness of the layer.

These quantities can be determined from the curve relating equivalent height h' and frequency f obtained by the usual method of vertical-incidence radio sounding. The ordinary ray critical frequency of the layer, f^0, can be read off this curve by inspection, while the parameters h_0 and y_m can be determined from the same curve by the method described in the Appendix to Part I of this paper.

Now one of the advantages of the solution previously given is that, by means of it, the relation between the oblique incidence critical frequency $f_{\max}$ (i.e. the maximum usable frequency) and the vertical-incidence critical frequency f^0 can be expressed as

$$f_{\max}/f^0 = x(h_0, y_m, D), \qquad \ldots\ldots(1)$$

where D is the distance of transmission, and where the function x represents what we may call the "maximum usable frequency factor" (M.U.F. factor) by which the experimentally determined quantity f^0 must be multiplied to give the required quantity $f_{\max}$. The special convenience of the use of (1) in practice arises from the fact that the variation of the function x throughout the hours of the day and the seasons of the year can be ascertained from the results of a preceding year (at any rate for quiet ionospheric conditions), whereas the fairly substantial changes of f^0 from day to day (and even from hour to hour) are not at present predictable.

The relation between the M.U.F. factor and the distance of transmission D for appropriate ranges of the parameters y_m and h_0 may be exhibited in two series of curves, samples of which are reproduced in small scale * in figures 1 and 2.

---

* A complete set of larger scale graphs will be supplied to *bona fide* users on application to the Superintendent, Radio Division, National Physical Laboratory, Teddington, Middlesex. The copies supplied, unlike the reproductions in this paper, have a graticule.

In figure 1, two sample sets of curves show the relation between the M.U.F. factor and D for different appropriate values of the height of maximum layer ionization $(y_m + h_0)$, the ratio y_m/h_0 being constant and equal to 0 and 0·2 respectively. Probably a more convenient representation is that of figure 2, in which the value of the M.U.F. factor is shown as a function of y_m/h_0 for different representative distances of transmission D and different values of $(y_m + h_0)$.

The wide range of information embodied in figures 1 and 2 should be noted. The complete set of curves covers a range of layer heights and thicknesses appropriate to the various ionospheric layers. The limiting curves for layers of zero thickness are identical, as they should be, with those drawn for sharp boundary reflection.

§3 SOME NOTES ON THE PRACTICAL DETERMINATION OF y_m AND h_0

In the Appendix to Part I, a graphical method of determining y_m and h_0 from (h', f) vertical incidence data was described. Briefly, this method consists in the graphical determination of the parameters y_m and h_0 in the equation representing the relation between equivalent height h' and frequency f, namely,

$$h' = h_0 + \frac{y_m}{2} \log_e \frac{f^0 + f}{f^0 - f}. \qquad \text{......(2)}$$

Now practical use of this method has led us to note two points of importance in its application. In the first place it often happens that the experimental (h', f) curve cannot be represented by (2) over the whole frequency range. In that case it is necessary to choose the constants of the equation so that it represents closely that part of the (h', f) curve which is most important for our purpose. Secondly, since the theory as so far described relates only to the case of a single deviating layer, it is necessary to consider the modifications necessary to allow for the influence of refraction in an ionized layer situated below that in which the waves are ultimately reflected. These two matters are considered in greater detail in sections (a) and (b) immediately below.

(a) *The importance of the high-frequency portion of the vertical incidence (h', f) curve*

For a plane ionosphere, the equivalent path of waves of frequency f, incident at an angle i_0, can readily be expressed in terms of the equivalent path of a frequency $f \cos i_0$ incident normally (Martyn, 1935). Now in Part I of this paper it was shown that a curved ionosphere behaves like a plane ionosphere if we assume that the vertical-incidence critical frequency is reduced to

$$\left(f^0 - \frac{f^2}{f^0} \cdot \frac{y_m}{R + h_0} \sin^2 i_0\right).$$

The equivalent vertical-incidence frequency in the curved ionosphere case, expressed as a fraction of the critical frequency, is therefore given approximately by

$$x_1 = \frac{f \cos i_0}{f^0\left(1 - \frac{f^2}{f^{02}} \frac{y_m}{R + h_0} \sin^2 i_0\right)}. \qquad \text{......(3)}$$

Actual calculations of maximum usable frequencies, using appropriate values of y_m and h_0, show that we are generally concerned with values of x_1 greater than 0·9. Hence it is of great importance to determine y_m and h_0 from that part of the vertical incidence (h', f) curve which indicates approach to penetration (i.e. the parabolic formula should express the distribution of ionization with height near the level of maximum, $(y_m + h_0)$). For this purpose the equivalent heights should be known accurately for frequencies between $0{\cdot}9f^0$ and f^0.

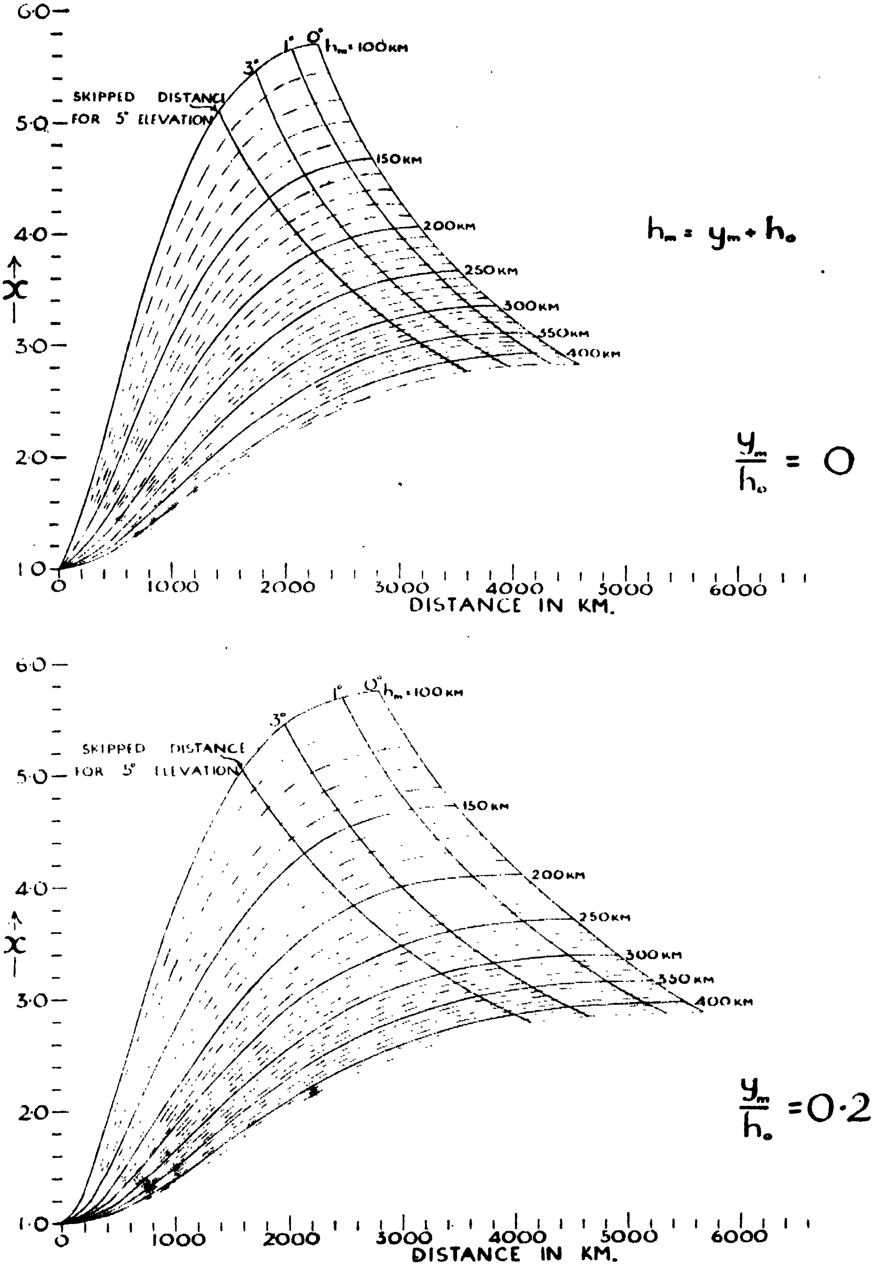


Figure 1. M.U.F. factor curves for a parabolic layer.

(b) *The influence of lower layers on the calculation of the M.U.F. factor*

Let us consider the effect of a refracting E-layer below the reflecting F_2 region, the earth being taken as flat. (The argument which follows can be equally well applied to the case of region F_1 below region F_2.) We assume that the electronic density distribution with height is of the form shown by the full curves in figure 3, both regions E and F_2 being of parabolic form. Let h, y and f^0, with appropriate suffixes, be the usual constants of the two regions. We consider the trajectory of

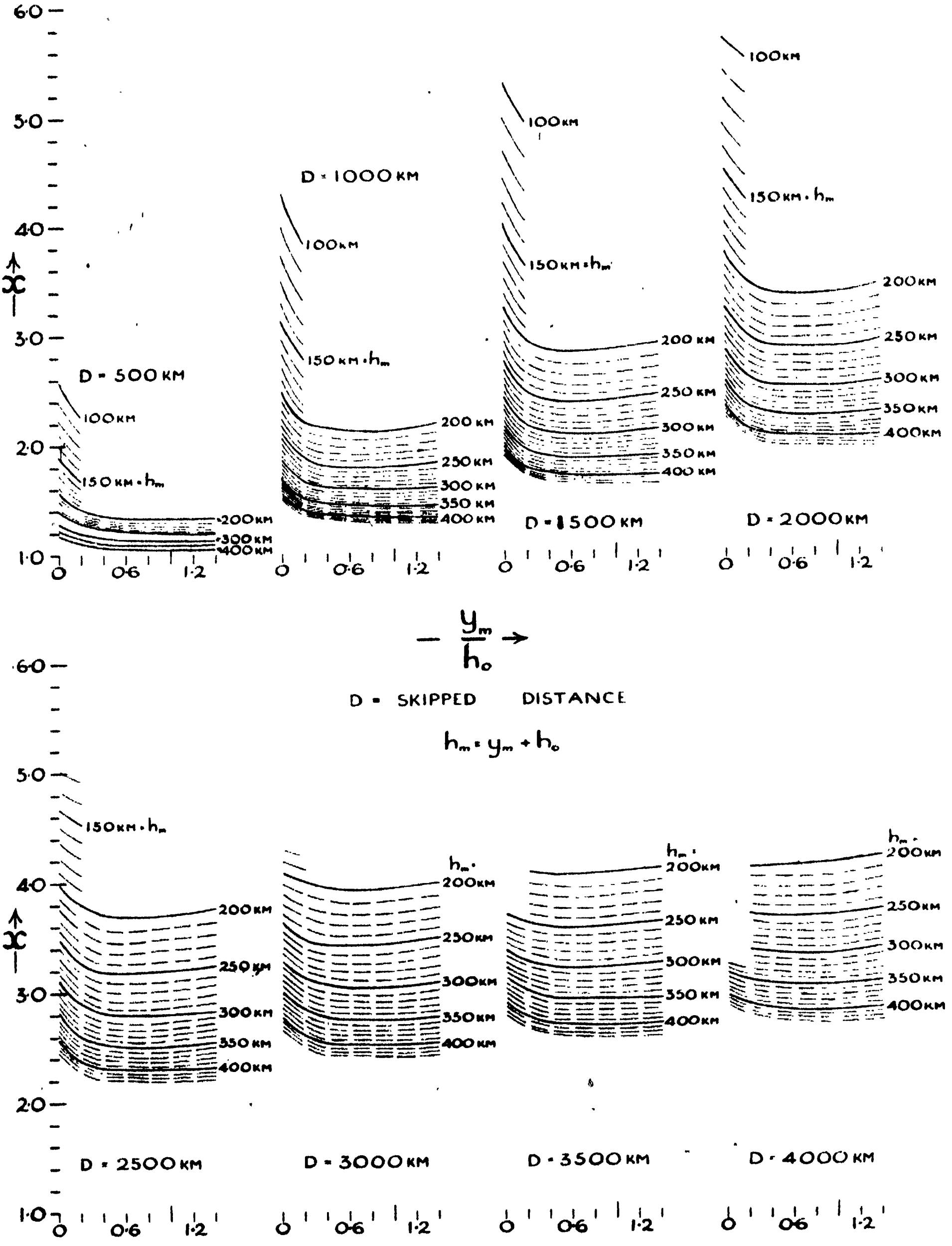


Figure 2. Cross-sections of M.U.F. factor curves at various distances.

waves of frequency f incident at an angle i_0 on the E-layer (figure 4). After passing through that layer they impinge on the F-layer at the same angle of incidence.

In the absence of the lower layer, the horizontal range would be given by

$$S'R' = y_F x_F \sin i_0 \log_e \frac{1 + x_F \cos i_0}{1 - x_F \cos i_0} + 2h_F \tan i_0, \qquad \ldots\ldots(4)$$

where $$x_F = f/f_F^0.$$

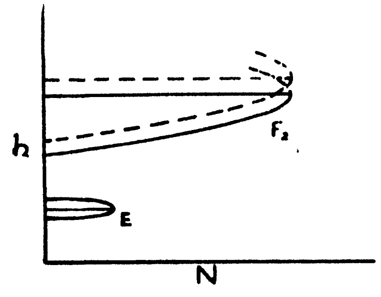


Figure 3.

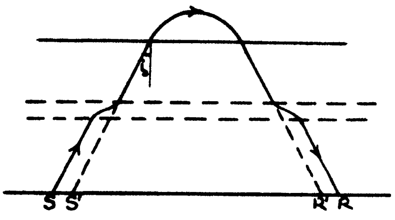


Figure 4.

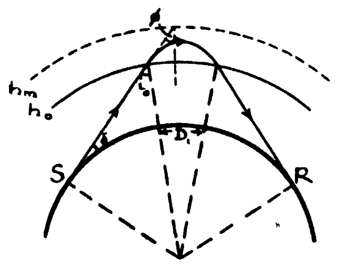


Figure 7.

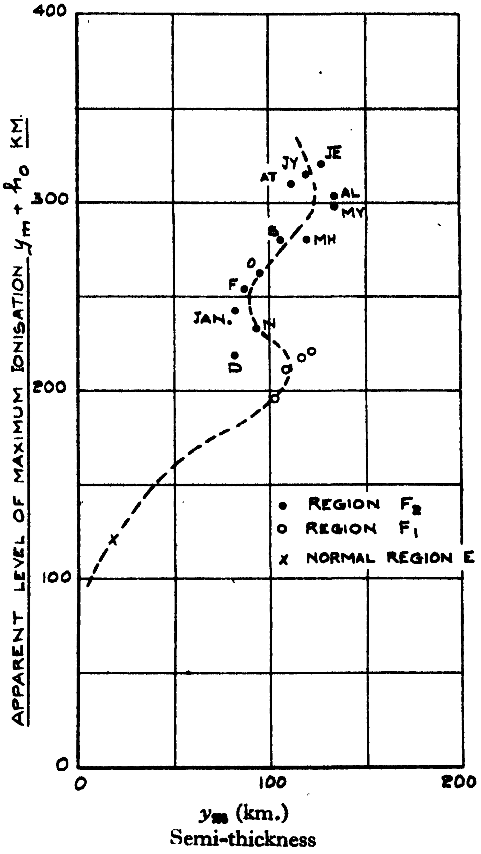


Figure 5. Mean variation of apparent semi-thickness y_m with apparent level of maximum ionization $y_m + h_0$. Monthly mean noon values based on Slough normal incidence data (1943).

With the group-retarding effect of the lower layer included, the actual range is

$$SR = x_F y_F \sin i_0 \log_e \frac{1 + x_F \cos i_0}{1 - x_F \cos i_0} + 2h_F \tan i_0$$
$$+ 2 \tan i_0 y_E \left\{ x_E \cos i_0 \log_e \frac{x_E \cos i_0 + 1}{x_E \cos i_0 - 1} - 1 \right\}, \qquad \ldots\ldots(5)$$

where $$x_E = f/f_E^0.$$

Now $$x_E \cos i_0 \log_e \frac{x_E \cos i_0 + 1}{x_E \cos i_0 - 1} > 1,$$

so that, for a given frequency incident at a given angle, the effect of a region beneath the main deviating region is to increase all distances, including "skip" distance; that is to say, the M.U.F. of the upper region for a given horizontal range will be decreased.

Consider now the M.U.F. calculation for these conditions. At vertical incidence the equivalent height for a frequency f is given by

$$h' = \frac{y_F}{2} x_F \log_e \frac{1+x_F}{1-x_F} + h_F - 2y_E + y_E x_E \log_e \frac{x_E+1}{x_E-1}. \qquad \ldots\ldots(6)$$

Expanding the second logarithmic term in (6), we have, since $f/f_E^0 = x_E > 1$,

$$\log_e \frac{f+f_E^0}{f-f_E^0} = \frac{2f_E^0}{f} + \frac{2}{3}\left(\frac{f_E^0}{f}\right)^3 + \ldots.$$

Hence $$h' \simeq \frac{y_F}{2} \cdot \frac{f}{f_F^0} \log_e \frac{f_F^0+f}{f_F^0-f} + h_F + \frac{2}{3} y_E \left(\frac{f_E^0}{f}\right)^2. \qquad \ldots\ldots(7)$$

Now in using the vertical incidence (h', f) curve to deduce the semi-thickness and height of lower boundary of the reflecting layer, we confine our measurements to frequencies greater than $0{\cdot}9f_F^0$. For such frequencies we can write $f \simeq f_F^0$, and the effect of the lower layer will be to add a term of approximate magnitude $\frac{2}{3}y_E(f_F^0/f_E^0)^2$ to the calculated value of the height of lower boundary (see dotted curve in figure 3). We have examined theoretically the extent to which the use of this fictitious height in a M.U.F. calculation, for both flat and curved earth cases, compensates for the effect of the lower layer,* and find that, for values of the layer constants characteristic of normal ionospheric conditions, the errors are rarely likely to exceed 4%.

§4. THE PRACTICAL USE OF THE GRAPHS

As has been explained above, the M.U.F. factor, for any distance of transmission, can be determined when the quantities y_m and h_0 are known. To get the most reliable estimate of the factor at any time, it is obviously desirable that y_m and h_0 should be determined from vertical-incidence measurements at a place midway between the sending and receiving stations. Such a course is, however, rarely practicable, and it is usually necessary to forecast the probable values of y_m and h_0 from past experience. As a result of a study made at the Radio Research Station, Slough (Lat. 51½° N.), the diurnal, seasonal and sunspot cycle variations of these quantities at this latitude are now known. An account of the results of this investigation will be given in a later communication.

As specimens of the range of values found, we may quote some results for region F_2 for the year 1944, a year near sunspot minimum. These are given in table 1 below.

It will be seen that the height of the level of maximum ionization $(y_m + h_0)$ usually reaches a minimum near noon, and that this level undergoes a seasonal change, being higher in summer than in winter.

---

* A note on the formulae, etc. involved in this examination is given in the Appendix.

Table 1

1944	G.M.T.	0000	0600	1200	1800
July	y_m+h_0 (km.) y_m/h_0	313 0·43	299 0·54	276 0·57	296 0·58
September	y_m+h_0 (km.) y_m/h_0 ..	334 0·43	260 0·56	263 0·58	286 0·54
December	y_m+h_0 (km.) y_m/h_0	335 0·54	297 0·52	241 0·47	274 0·59

A further sample of values is shown in figure 5, where the monthly mean noon values of y_m for the whole of the year 1943 are plotted against the corresponding values of y_m+h_0 obtained from Slough data. Most of the plotted points refer to region F_2, but some values for region E and region F_1 are also included. The individual estimates of y_m and (y_m+h_0) often show marked variability, but the mean points shown on the diagram indicate clearly the general tendency for larger values of y_m to be associated with larger values of (y_m+h_0). This general tendency is characteristic of measurements of y_m and h_0 made at all times of day and night, but is particularly well defined in the noon values. The physical significance of this increase of thickness of the ionized layer with increasing height was discussed by one of us some years ago (Appleton, 1939).

Calculations of y_m and h_0 from samples of experimental normal-incidence (h', f) curves observed at several widely separated ionospheric observatories indicate that the mean variation of y_m with (y_m+h_0) shown in figure 5 is most probably indicative of that which occurs over a wide range of latitude and longitude.

§5. ESTIMATION OF M.U.F. DIRECTLY FROM THE NORMAL INCIDENCE (h',f) CURVE

When the two parameters y_m and h_0 are known, the M.U.F. factor, and hence the M.U.F. itself, can be read directly from the set of standard curves. On the other hand, if a standard type of (h',f) record is always available, then it is more convenient to be able to read the M.U.F. directly by applying some other set of curves to the experimental (h',f) curve. Curves of this type, known as "transmission curves", were first described by Smith (1937) and by Millington (1938), and have formed the basis of M.U.F. calculations at a number of ionospheric observatories. In the case of the "parabolic layer" method, it is an extremely simple matter to convert the M.U.F. factor curves described above to a form suitable for direct application to any experimental normal incidence (h',f) curves. The transmission curves corresponding to any specified distance of transmission consist in the envelopes of a series of theoretical normal incidence (h', f) curves drawn for parabolic layers having the constants y_m and h_0 related by the mean curve shown in figure 5. Sets of transmission curves constructed in this way are shown in figure 6.

We have already mentioned that individual estimates of y_m and (y_m+h_0) often show considerable divergence from the mean curve given in figure 5, and it might be expected that the assumption of this fixed relation between these two quantities would result in appreciable errors in the estimates of the M.U.F. Two factors, however, tend to make such errors of quite small magnitude. In the first place, as we have already shown, the M.U.F. factor is very largely determined by the magnitude of $(y_m + h_0)$ and depends only to a smaller degree on the actual magnitude of the semi-thickness y_m. An examination of the standard M.U.F. curves in conjunction with our knowledge of the actual variations of y_m, has shown that, as far as M.U.F. calculations are concerned, the effect of fluctuations about the values given by the mean curve shown in figure 5 will be small. Secondly, the transmission-curve technique is largely self-compensatory with respect to variations in thickness of the layer, the result being that even the magnitudes of possible errors due to assuming a fixed law of variation of thickness with height are, in practice, further reduced.

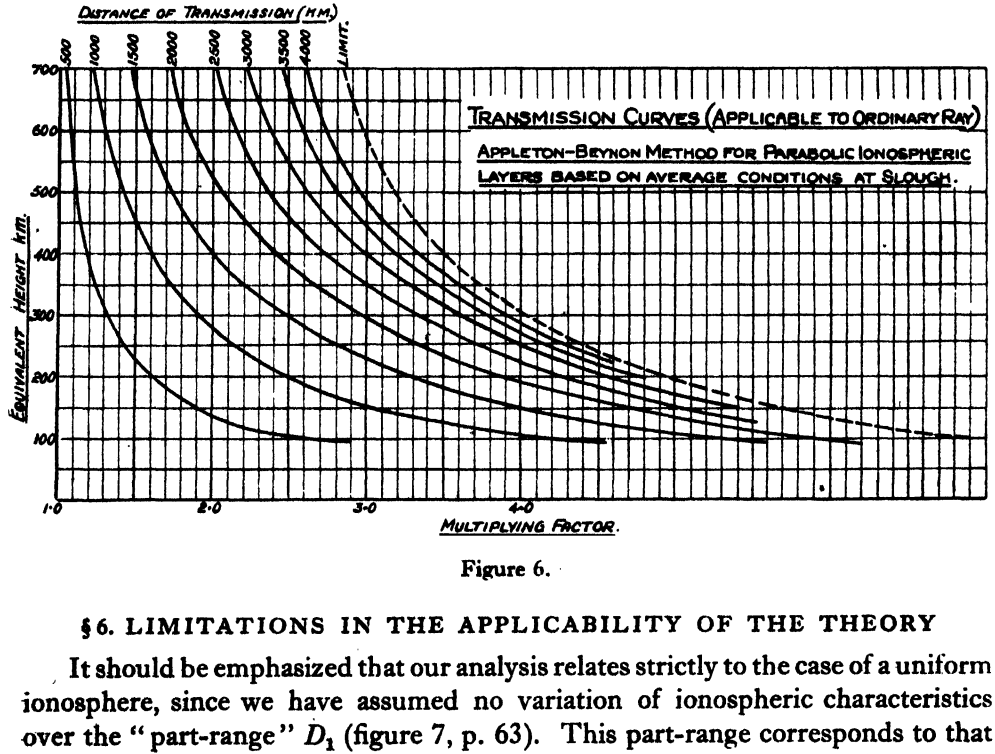


Figure 6.

§6. LIMITATIONS IN THE APPLICABILITY OF THE THEORY

It should be emphasized that our analysis relates strictly to the case of a uniform ionosphere, since we have assumed no variation of ionospheric characteristics over the "part-range" D_1 (figure 7, p. 63). This part-range corresponds to that portion of the transmission path which lies within the ionosphere itself. From the typical numerical calculation given in Part I it will be seen that, for a total range of 3500 km., D_1 may amount to as much as 1300 km. While it is probably correct, under most conditions, to consider the values of f^0, h_0 and y_m at the mid-point of this part-range as representative of the whole of the ionospheric track, we should expect that the results would be less reliable when sunset or sunrise conditions obtain over this section of the ionosphere. Under such non-uniform conditions lateral deviation from the great circle path may also occur for certain directions of transmission; in addition, asymmetrical ray tracks are to be expected.

Still greater difficulties are encountered when it is attempted to apply the relatively simple theory illustrated above to a case of long-distance transmission in which it is necessary to assume multiple hops. In such cases the changing ionospheric conditions will result in the distance between successive earth reflections being unequal. In such circumstances the application of the above analysis becomes a complicated matter.

§7. THE MAXIMUM FACTOR AT A GIVEN ANGLE OF INCIDENCE

For certain problems, particularly for extreme distance transmission, it may be useful to know the maximum radio-frequency which can be returned from the ionosphere at any specified angle of incidence, irrespective of the horizontal distance involved. Referring to figure 7, the part-range D_1, for a frequency f incident at an angle i_0, is given by

$$D_1 = \frac{R}{R+h_0} \sin i_0 \, . \, x \, . \, y_m \log_e \left\{ \frac{1 - \dfrac{x^2 y_m}{R+h_0} \sin^2 i_0 + x \cos i_0}{1 - \dfrac{x^2 y_m}{R+h_0} \sin^2 i_0 - x \cos i_0} \right\}. \qquad \ldots\ldots(8)$$

For finite values of D_1, it is clear that x must satisfy the inequality

$$\frac{x \cos i_0}{1 - x^2 \dfrac{y_m}{R+h_0} \sin^2 i_0} \leqslant 1. \qquad \ldots\ldots(9)$$

To terms in $\dfrac{1}{R+h_0}$ equation (9) can be written in the form

$$x^2 \left[1 - \sin^2 i_0 \left(\frac{R+h_0}{R+h_0+y_m} \right)^2 \right] \leqslant 1 \qquad \ldots\ldots(10)$$

or

$$x \leqslant \sec \phi, \qquad \ldots\ldots(11)$$

where ϕ is the angle shown in figure 7. Equation (9) can also be reduced to the condition

$$x \leqslant \left\{ \frac{\sqrt{1 + \dfrac{4y_m}{R+h_0} \tan^2 i_0} - 1}{\dfrac{2y_m}{R+h_0} \tan^2 i_0} \right\} \sec i_0. \qquad \ldots\ldots(12)$$

This inequality determines the absolute maximum value of x for any given angle of incidence. It will be noted that for any fixed angle of incidence this absolute maximum factor is determined by the value of the ratio $y_m/(R+h_0)$. Since $h_0 \leqslant R$, an approximate form of (12) is

$$x \leqslant \left\{ \frac{\sqrt{1 + \dfrac{4y_m}{R} \tan^2 i_0} - 1}{\dfrac{2y_m}{R} \tan^2 i_0} \right\} \sec i_0. \qquad \ldots\ldots(13)$$

Thus, for any given angle of incidence, the maximum value of the multiplying factor x depends principally on y_m, the semi-thickness of the reflecting region. It may be noted that when $R \to \infty$, equation (12) reduces, as it should, to the usual expression for a plane ionosphere, namely $x \leqslant \sec i_0$.

It is often, however, more convenient to express this limiting factor in terms of the angle of emission δ relative to the ground at the sender (see figure 7).

Now

$$\tan^2 i_0 = \frac{R\cos^2\delta}{R\sin^2\delta + 2h_0}, \qquad \ldots\ldots(14)$$

approximately, so that (12) becomes

$$x \leqslant \left\{ \frac{\sqrt{1 + \dfrac{4y_m}{R+h_0} \cdot \dfrac{R\cos^2\delta}{R\sin^2\delta + 2h_0}} - 1}{\dfrac{2y_m}{R+h_0}\left(\dfrac{R\cos^2\delta}{R\sin^2\delta + 2h_0}\right)} \right\} \sqrt{\frac{R+2h_0}{R\sin^2\delta + 2h_0}}. \qquad \ldots\ldots(15)$$

For the tangential ray $\delta = 0$, and expression (15) becomes

$$x \leqslant \left\{ \frac{\sqrt{1 + \dfrac{4y_m}{R+h_0} \cdot \dfrac{R}{2h_0}} - 1}{\dfrac{y_m}{R+h_0} \cdot \dfrac{R}{h_0}} \right\} \sqrt{\frac{R+2h_0}{2h_0}}. \qquad \ldots\ldots(16)$$

Since $h_0 \ll R$, an approximate expression for the case $\delta = 0$ is

$$x \leqslant \sqrt{R}\left\{ \frac{\sqrt{2h_0 + 4y_m} - \sqrt{2h_0}}{2y_m} \right\}. \qquad \ldots\ldots(17)$$

Figures 8, 9 and 10 have been drawn to illustrate formulae (11), (13) and (14) for specified values of the layer constants y_m and h_0. From these graphs the limiting factors at any given angle of elevation can easily be determined.

§ 8. COMPARISON OF PRACTICAL RESULTS WITH THEORETICAL PREDICTIONS

The above method of calculating the M.U.F. as a function of distance, for specified ionospheric conditions, has formed the basis of all the M.U.F. predictions, for different parts of the world, which have been issued confidentially during the war to British Service and Civil users from the Radio Research Station, Slough. We have therefore had considerable experience of the use of the method and so have been able to check its general applicability.

As has been seen above, the M.U.F. is calculated by multiplying the vertical-incidence critical frequency by a certain M.U.F. factor. Now, generally, the critical frequency of the *F* layer exceeds that of the *E* layer, but, on the other hand, the M.U.F. factor for a given value of distance *D* is greater for the *E* layer than for the *F* layer. We cannot, therefore, say that the one or the other of the two reflecting layers always determines the M.U.F. Actually it turns out that, in winter daytime, the *F*-layer conditions control the M.U.F., whereas in summer daytime the normal *E* layer is in control. The recombination of the ionization in the *E* layer is always so rapid after sunset that practically throughout the year the night-time M.U.F. is controlled by the *F* layer.*

The success of the application of the method in ionospheric forecasting evidently depends on an accurate forecast of ionospheric conditions at the mid-point of the overhead-ray trajectory. The forecasting of normal *E*-layer conditions is accurate to a fairly high degree, so regular is this layer in its behaviour. Apart

---

* Exceptions to this general rule sometimes occur during a summer night when abnormal *E* ionization may control the M.U.F. (*vide infra*, § 8 (c)).

from the effects of ionospheric storms, the forecasts of *F*-layer conditions, though not so reliable as in the case of the normal *E* layer, are sufficiently accurate to permit the M.U.F. forecasts to be reliable enough for practical usefulness. On the other hand, as we shall show later, the incidence of abnormal *E*-layer ionization very often permits satisfactory communication on frequencies exceeding the predicted values based on estimates of the normal *E*- and *F*-layer ionospheric conditions. Such events are found to occur most frequently in daytime during summer when, as is well known, abnormal *E*-layer ionization is most frequently experienced.

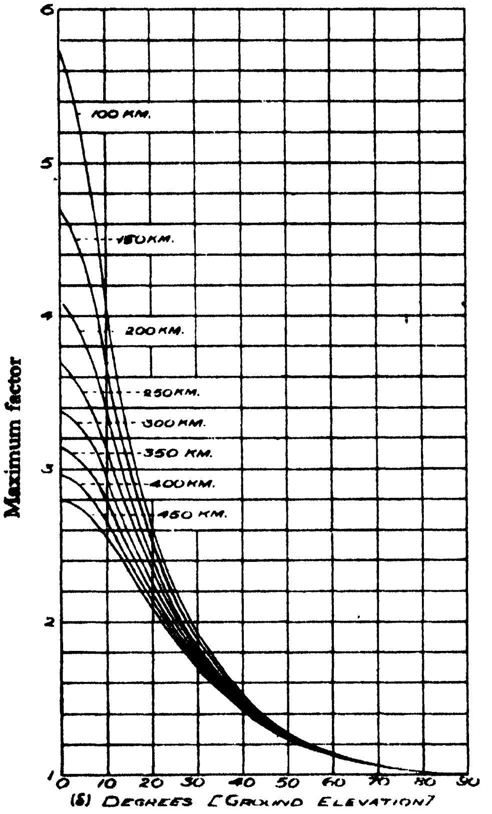


Figure 8. Maximum factor for different values of *hm*.

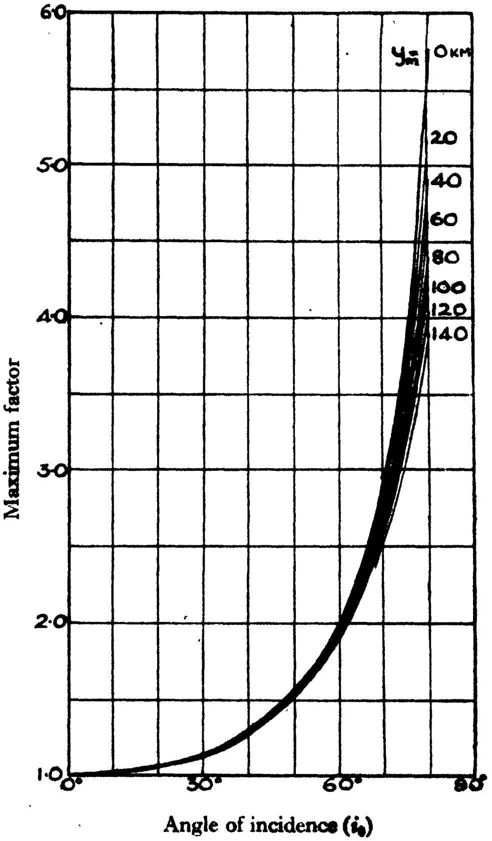


Figure 9.

Before proceeding to discuss the cases of abnormal *E*-layer transmission by way of illustrations from practical communication data, we turn to consider a number of points of physical interest in connection with the practical realization of a M.U.F. under normal conditions, such as occur when the *F* layer controls its value.

(a) *Signal intensity phenomena associated with the* M.U.F.

In making check measurements of predicted M.U.F. values from signal-intensity records of short-wave transmitting stations we have noted interesting interference phenomena as the M.U.F. conditions are approached, culminating in a very high signal value at, or near, the M.U.F. itself. All such phenomena appear to be satisfactorily explained in terms of the parabolic-layer theory of refraction outlined in this paper. For convenience we shall discuss one or two theoretical points first. We shall take the case of a flat earth since it is simpler than the curved-earth treatment and illustrates all the essential physical phenomena.

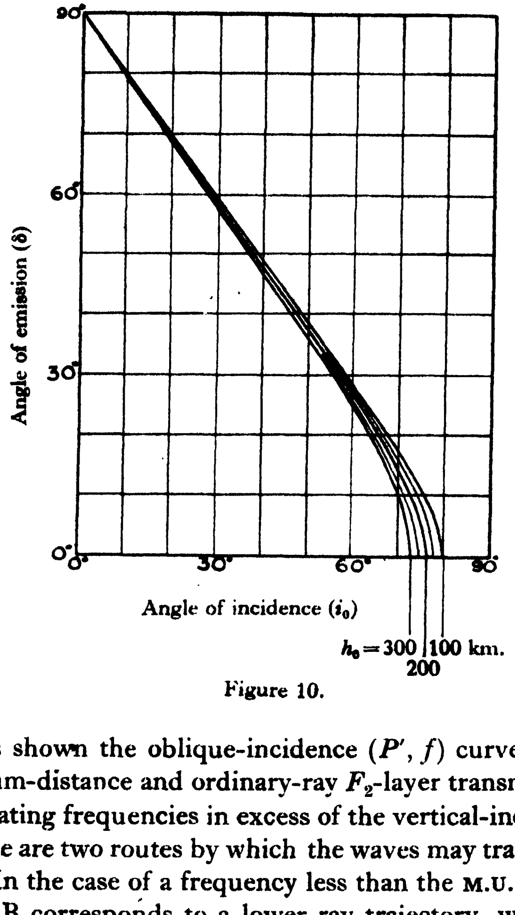


Figure 10.

In figure 11 is shown the oblique-incidence (P', f) curve calculated by our theory for a medium-distance and ordinary-ray F_2-layer transmission. It will be seen that, for operating frequencies in excess of the vertical-incidence penetration frequency $f^0_{F_2}$ there are two routes by which the waves may travel from the sender to the receiver. In the case of a frequency less than the M.U.F., say f_1, the lower equivalent path AB corresponds to a lower-ray trajectory, while the equivalent path AC corresponds to an upper-ray * trajectory. The reception of two sets of waves at the receiver causes interference phenomena if the two amplitudes are comparable.† This is especially the case as f_1 approaches the M.U.F. Now the state of interference as regards the phase difference of the two interfering components is determined by the difference of the optical paths of the upper and lower rays, and it can be shown that this quantity, for a frequency f_1, is equal to the area BNC divided by the frequency f_1. Thus, as f_1 approaches the M.U.F., the area BNC decreases and interference maxima and minima in signal strength are experienced culminating in a pronounced maximum at, or near, the exact M.U.F. value. Actually it is easy to show, by the usual method of ray tracing, that when f is equal to the M.U.F. the receiving station is situated on a caustic. The interference

---

* This upper ray is often termed the Pedersen ray. The upper and lower rays have, of course, different angles of incidence on the layer, the upper ray having the lesser angle of incidence.

† That maximum and minimum of signal amplitude are experienced just *outside* the skip zone has been noticed by Grosskopf (1940). He explains these phenomena as due to interference between the Pedersen ray and the normal ray, but gives no quantitative treatment of them.

phenomena experienced at the boundary of the skipped distance have therefore a certain resemblance to those experienced in the case of the rainbow. Just within the illuminated area there are the maxima and minima of intensity described above, and interpreted physically as due to the simultaneous reception of the lower and upper rays, while the intensity within the shadow (" silent zone ") falls off exponentially within the skipped distance.

In practice, the high-angle Pedersen rays are often heavily attenuated except at frequencies near the M.U.F. Furthermore, under suitable conditions, both the ordinary and extraordinary ray components may be present. A practical oblique-incidence (P', f) record might thus be of the form shown in figure 12, and, in this case, interference effects between two, three or even four rays will occur. Thus, referring to figure 12, as the frequency f_1 is increased to $f_1 + \delta f_1$ there will be a small increase in optical path difference between the two magneto-ionic components given by the area ABCD/f_1. Similarly, for a change from f_2 to $f_2 + \delta f_2$, the changes in optical path difference, (*a*) between the two ordinary ray components (upper and lower trajectories) and (*b*) between the low-angle ordinary and extraordinary components, will be proportional to the areas EFGH and HGLM respectively. Hence, for conditions such as those represented by figure 12, we may expect the following sequence of events as the frequency in use steadily approaches the M.U.F.:

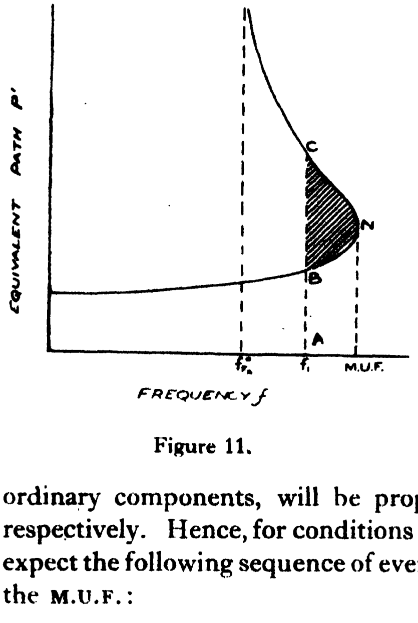


Figure 11.

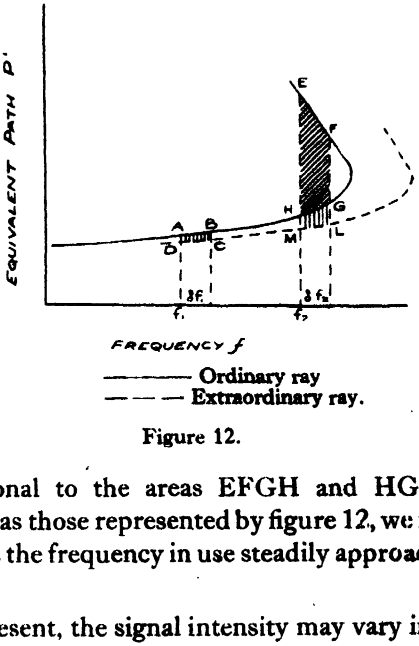


Figure 12.

(i) At first, with only a single ray present, the signal intensity may vary in an arbitrary random manner.

(ii) With the growth in amplitude of the extraordinary ray we may expect rhythmic fading of slow period, coupled with equally regular polarization changes. As the optical path difference between these two components gradually increases more rapidly (see figure 12), this rate of fading should slowly increase.

(iii) When the effect of the high angle (Pedersen) ordinary ray becomes appreciable, a rapid fading should be superposed on the slow fading.

(iv) As the M.U.F. of the ordinary component is approached, the ordinary-extraordinary fading becomes more rapid and the upper-lower ordinary ray fading slower, until the two rates may be of comparable magnitude.

(v). After the penetration of the ordinary component we are left with fairly rapid fading between the two extraordinary rays (upper and lower), this rate then decreasing to zero at the M.U.F. for the extraordinary ray.

In practice it is not readily feasible to study these phenomena either by studying the signal intensity as the frequency is steadily increased up to, and beyond, the M.U.F. or, in the fixed-frequency case, by studying the signal intensity variations in space round the edge of the skipped distance. But under conditions of steadily increasing (or decreasing) ionization such as occurs at sunrise (or sunset), when the changing M.U.F. value passes through the fixed operating frequency used, the varying features of the interference system described above appear as a sequence in time.

For example, in measurements made on Oslo in the evening of 25 January 1946, when the *F*-layer ionization was decreasing (which would simulate the conditions illustrated in figure 12 as f_1 approached, equalled, and then exceeded the M.U.F.), the signal intensity record shown in figure 13 was obtained.

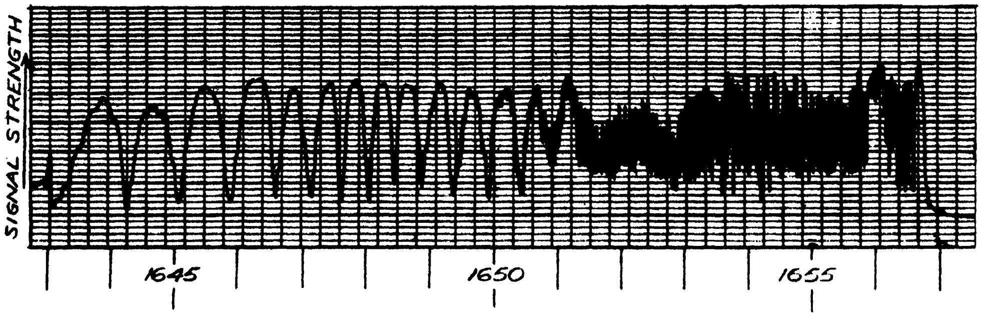


Figure 13. Variation in signal strength near M.U.F. Signal from Oslo received at Slough. Frequency 9·53 Mc./s. Distance 1140 km. 25 Jan. 1946.

Prior to 1643·5 G.M.T. the received signal intensity varied slowly in a random manner. At this time the slow rhythmic fading commenced at the rate of about 1·3 c./min. Within five minutes this rate of fading increased to about two per minute. At 1649 G.M.T. we note the first signs of the rapid fading due to the reception of the Pedersen ray; this increased in amplitude until by 1651 G.M.T. we get fast fading at the rate of about 37 cycles per minute. This then persists until the disappearance of the ordinary component at 1656 G.M.T., after which slower fading (18 c./min.), due to interference between the two extraordinary components, continues for about 40 seconds before the signal finally decreases to zero just before 1657 G.M.T. The high peaks in signal intensity which occur exactly at the maximum usable frequencies (1656 and 1656·7 G.M.T.) do not show up well on this record because of overloading of the receiver for very strong input signals. It is interesting to compare the observed rates of fading shown on this record with values calculated for a parabolic layer from normal-incidence (h', f) observations made at Slough and Burghead at the time of the field-strength record. (The mean latitude of the Slough-Oslo transmission path is approximately equal to the mean latitude of Slough and Burghead.) These calculated values are given in table 2 below.

Table 2

Frequency f (Mc./s.)	Ratio $\frac{f}{\text{M.U.F.}}$	Difference in optical path (δP)		Calculated rate of fading (cycles per minute)	
		Pedersen-low ray (km.)	Ordinary-extraordinary (km.)	Pedersen-low ray	Ordinary-extraordinary
9·53	0·91	18·0	0·1	170	1·0
	0·94	8·0	0·15	110	2·1
	0·97	2·2	0·20	60	5·5
	0·985	0·7	0·38	36	20·0

It will be seen that the calculated rates of fading agree reasonably well with the experimental values.

(b) *The variability of the F_2 layer critical frequency*

In applying normal incidence ionospheric data to communication problems it is often necessary to estimate conditions at a point many hundreds of kilometres from the nearest ionospheric laboratory. There is ample evidence to show that for small latitude changes (up to, say, 10°) the variations in the M.U.F. factor are small compared with those in the normal incidence critical frequency of the F_2 layer (there is no such variability in the case of the E-layer critical frequency). Forecasts have, of course, to be made using monthly mean values and, even under quiet ionospheric conditions, we have found that the standard deviation from the monthly mean value may be of the order of 12 to 20%. (An extended analysis of this phenomenon made by Appleton and Naismith has shown that, for the latitude of Slough (51½° N.) and for years of marked solar activity, the spread of the daily values is generally greatest for the equinox periods March and September, indicating clearly the influence of magnetic storms.) Thus interpolated or extrapolated values of the M.U.F. based on monthly mean values will not necessarily be really accurate for any given day. To allow for these variations it is customary, for practical purposes, to advocate operation on frequencies 15% less than the predicted value based on monthly mean values of the normal incidence critical frequency.

(c) *The influence of abnormal E-layer ionization on oblique incidence transmission*

There is now available a considerable body of evidence showing that satisfactory radio communication can quite frequently be established on frequencies higher than the maximum usable frequencies predicted by the method described above, which is based on the regular variations of the E and F_2 layers. During the war many examples of this phenomenon were brought to our attention by the British Broadcasting Corporation Engineering Division, who cited to us cases of satisfactory long-distance broadcasting on frequencies which considerably exceeded the maximum usable frequency predicted in the normal way. The information kindly furnished to us by the B.B.C. related both to "multiple-hop" and "single-hop" transmission. In both cases it had been noted that discrepancies of the kind in question occurred most frequently in daytime and in summer.

Now as far back as 1935 attention was drawn to the importance of abnormal *E*-layer ionization in facilitating long-distance high-frequency communication. Appleton and Naismith (1935), for example, found that this type of ionization, particularly in summer daytime, often returns radio energy copiously at frequencies greatly exceeding the critical frequency of the normal *E* layer. "Under such conditions", they concluded, "the upper limit of frequency usable in long-distance communication can be abnormally high, since the usual restriction by F_2 region electron limitation is then not operative".

We have made a special study of the B.B.C. "single-hop" transmissions which relate to the reception of Daventry short-wave signals in Gibraltar and Algiers, the results of which appeared in a confidential paper circulated by the Radio Research Board in August 1944. Generally we found that the B.B.C. results could be explained partly by reflections from the normal *E* and F_2* layers and especially by the intervention of abnormal *E*-layer ionization which provides copious reflection in summer daytime.

More recently, some particularly striking evidence relating to communication on frequencies exceeding the predicted values has been brought to our notice by Dr. L. P. Wheeler, Chief of the Technical Information Division of the Engineering Department of the Federal Communications Commission, U.S.A., to whom we are indebted for permission to quote here. During the summer of 1944 Dr. Wheeler found that strong signals were received in the 40 to 50 Mc./s. band at distances of the order of 1000 to 1500 km. from the sender, whereas for the particular conditions in question the maximum usable frequency for reflection by the F_2 layer or by the normal *E* layer would have been about 12 and 16 Mc./s. respectively. It is clear, therefore, that abnormal *E*-layer ionization was responsible as the reflecting stratum.

Now the M.U.F. factor for the abnormal *E* layer for a distance of 1500 km. is approximately 5·0, so that for the successful transmission of, say, 44·3 Mc./s. the critical penetration frequency of the layer must have exceeded 8·9 Mc./s. So high a critical frequency is only characteristic of what Appleton and Naismith have termed the "intense *E*-layer" condition under which the penetration frequency of the abnormal *E* layer exceeds the normal F_2 layer critical frequency.

To test whether an explanation along these lines was the correct one, we have compared the reception results of Dr. Wheeler with the information available concerning the incidence of "intense *E*-layer" ionization in the U.S.A. In figure 14 are shown plotted the monthly figures of the time of occurrence of substantial reflection of 44·3 Mc./s. waves over a distance of 1428 km., together with the frequency of occurrence of normal-incidence critical frequencies of the abnormal *E* layer exceeding 9 Mc./s. at vertical incidence at Washington. In each case the observations refer to the period 0600–2400 L.M.T. There can be little doubt, from a comparison of these graphs, that the phenomena are related. An exact and detailed correspondence would not be expected in view of the fact that the observations relate to different sites, and in view of the well

---

* Allowance for the "longitude effect" described recently by one of us (Appleton, 1946) turned out to be particularly important in this connection.

known non-uniform horizontal distribution of both abnormal *E*-layer and "intense *E*-layer" ionization. We can, however, be quite confident that these occasional successful transmissions on frequencies up to twice the normally predicted values are due to the intervention of abnormally dense sporadic ionization at the level of the *E* layer.

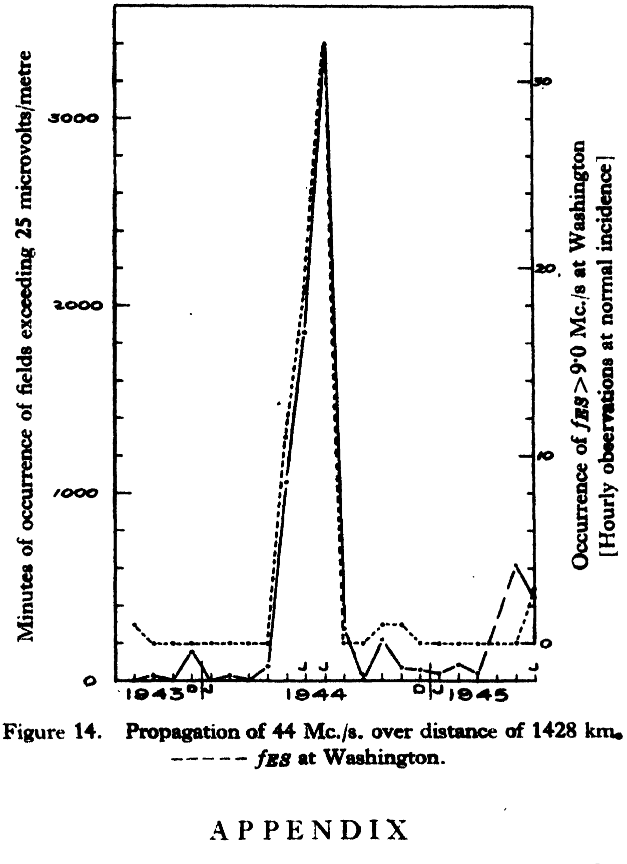


Figure 14. **Propagation of 44 Mc./s. over distance of 1428 km.**
------ f_{ES} at Washington.

APPENDIX

Note on the calculation of skipped distance for a compound ionospheric structure

For a plane ionosphere structure of the form shown in figures 3 and 4, the relation between i_0 and x for minimum D can readily be shown to be

$$\tan^2 i_0 \left\{ \frac{y_F x_F'^2}{1 - x_F'^2} - \frac{2 y_E x_E'^2}{x_E'^2 - 1} - (h_F - 2y_E) \right\}$$

$$= \left\{ (h_F - 2y_E) + x_E' y_E \log_e \frac{x_E' + 1}{x_E' - 1} + \frac{x_F' y_F}{2} \log_e \frac{1 + x_F'}{1 - x_F'} \right\}, \qquad \ldots\ldots(18)$$

where

$$x_E' = \frac{f_{\max}}{f_E^0} \cos i_0 \quad \text{and} \quad x_F' = \frac{f_{\max}}{f_F^0} \cos i_0.$$

For a curved ionosphere of the same structure, the angles of incidence of the waves on the *E* and *F* layers are unequal. In this case the corresponding formula is similar to the above, but with additional terms in $1/R$. Such a formula cannot, however, readily be put in an explicit form suitable for direct computation. Hence, in making exact M.U.F. calculations for comparison with the results of the approximate method outlined in the main text, it becomes necessary to determine the skipped distances directly by plotting distances against angle of incidence for given ionospheric conditions. For the curved ionosphere structure shown in figure 7 the true range D is given by

$$\left.\begin{aligned} D = {} & \frac{4R}{R+h_E} \,.\, x_E \,,\, y_E \,.\, \sin i_E \coth^{-1} \frac{x_E'}{1-\dfrac{x_E^2 y_E \sin^2 i_E}{R+h_E}} \\ & + \frac{2R}{R+h_F} \,.\, x_F \,.\, y_F \,.\, \sin i_F \tanh^{-1} \frac{x_F'}{1-\dfrac{x_F^2 y_F \sin^2 i_F}{R+h_F}} + D_2, \end{aligned}\right\} \quad \ldots\ldots(19)$$

where $D_2 = 2R(\theta_1+\theta_2)$, θ_1 and θ_2 being given in the equations

$$\tan i_E = \frac{R\sin\theta_1}{h_E + R(1-\cos\theta_1)}, \qquad \tan i_F = \frac{(R+h_E+2y_E)\sin\theta_2}{h_F - \{(R+h_E+2y_E)\cos\theta_2 - R\}}.$$

For the approximately equivalent region we should use the following formula for calculating i_0:

$$\tan^2 i_0 \left\{\frac{y_F x_F'^2}{1-x_F'^2} - (h_F+p)\right\} = \left\{h_F + p + \frac{x_F' y_F}{2} \log_e \frac{1+x_F'}{1-x_F'}\right\}, \qquad \ldots\ldots(20)$$

where $$p \simeq \tfrac{2}{3}\, y_E (f_E^0/f_F^0)^2.$$

The corresponding skipped distances would then be calculated in the usual way from

$$D = \frac{2R}{R+h_F+p} \,.\, x_F \,.\, y_F \,.\, \sin i_0 \tanh^{-1} \frac{x_F \cos i_0}{1-\dfrac{x_F^2 y_F \sin^2 i_0}{R+h_F+p}} + D_2', \qquad \ldots\ldots(21)$$

where $D_2' = 2R\theta'$ and θ' is given in the equation

$$\tan i_0 = \frac{R\sin\theta'}{h_F + p + R(1-\cos\theta')}.$$

ACKNOWLEDGMENT

The work described above was carried out as part of the programme of the Radio Research Board, and this paper is published by permission of the Department of Scientific and Industrial Research.

REFERENCES

APPLETON, E. V., 1939. "The Structure of the Atmosphere as deduced from Ionospheric Observations." *Quart. J.R. Met. Soc.*, **65**, 324.

APPLETON, E. V., 1946. "Two Anomalies in the Ionosphere." *Nature, Lond.*, **157**, 691.

APPLETON, E. V. and BEYNON, W. J. G., 1940. "The Application of Ionospheric Data to Radio Communication Problems", Part I. *Proc. Phys. Soc.*, **52**, 518.

APPLETON, E. V. and NAISMITH, R., 1935. "Some Further Measurements of Upper Atmospheric Ionization." *Proc. Roy. Soc.*, A, 871, **150**, 685.

GROSSKOPF, J., 1940. *Telegr. Fernspr.-Funk.- u. Fernsehtech.*, **29**, 127.

MARTYN, D. F., 1935. "The Propagation of Medium Radio Waves in the Ionosphere." *Proc. Phys. Soc.*, **47**, 323.

MILLINGTON, G., 1938. "The Relation between Ionospheric Transmission Phenomena at Oblique Incidence and those at Vertical Incidence." *Proc. Phys. Soc.*, **50**, 801.

SMITH, N., 1937. "The Extension of Normal Incidence Ionosphere Measurements to Oblique Incidence Radio Transmission." *J. Res. Nat. Bur. Stand., Wash.*, **19**, 89.

AN INFRA-RED SPECTROSCOPE WITH CATHODE-RAY PRESENTATION

By E. F. DALY and G. B. B. M. SUTHERLAND
Pembroke College, Cambridge

MS. received 30 *August* 1946

ABSTRACT. An instrument developed for the visual presentation of infra-red spectra in a cathode-ray screen is described. The radiation detector employed is a high-speed bolometer made by the Bell Telephone Laboratories having a time constant of less than 0·01 second. The performance of the instrument is such that a range of 2·5 to 3·5 μ, anywhere between 1 and 16 μ, can be scanned in 14 seconds. The cathode-ray screen used has very long persistence, so that a steady " picture " is obtained. The resolving power at these speeds is sufficiently good for most work on complex molecules. Illustrations of typical spectra are given and proposals are outlined for further improvement and development.

§1. INTRODUCTION

In recent years it has become very important to improve the speed with which absorption spectra may be plotted in the infra-red. The aim has been to make the recording of spectra a routine matter so that attention can be concentrated on the correlation and interpretation of data on series of molecules of related constitution. In this way bands which characterize particular chemical bonds and groupings can be identified with certainty and used as guides in elucidating the structure of molecules of unknown constitution. Several recording spectrometers have been described which plot absorption spectra between 2 and 15 μ (with moderate resolving power) in a time of the order of half an hour. (For example, Barnes *et al.*, 1945; Brattain and Beeck, 1942; McAllister *et al.*, 1941; Sutherland and Thompson, 1945; Wright, 1941.) However, in order to exploit fully the infra-red method of analysis, it is desirable to be able to see the infra-red spectrum, as this enables one to apply the method to problems in which the spectrum is changing rapidly, e.g. in chemical reactions, in changes of state and in other transitory phenomena. Quite apart from making these new fields accessible, such an infra-red spectroscope has great advantages as an inspection instrument and for doing a quick survey of a problem in order to enable one to decide whether a more detailed study is worth while, using a conventional recording instrument.

Before the instrument described here was commenced there had been only one attempt (Baker and Robb, 1943) to make an infra-red spectroscope using cathode-ray presentation. The principal controlling factor in such an instrument is the speed of the detector system. In the apparatus of Baker and Robb this consisted of a bolometer and a Moll micro-galvanometer, having a combined time constant of 0·2 sec. With this they could arrange to display 75 points in a spectrum in a time of one minute, i.e. about 1 point a second. The resulting picture must either cover an inconveniently small portion of the infra-red spectrum rather slowly, or a reasonable range with unsatisfactory resolving power. Unfortunately no absorption spectra were given by Baker and Robb from which the actual performance could be judged. In our instrument we employ as radiation detector a Bell Telephone Laboratories thermistor bolometer with a time constant of just under 0·01 sec. (Brattain and Becker, 1946). It is operated by radiation interrupted at 20 c./s. and is followed by an amplifier responding to the 20 c./s. output voltage of the detector, but not to slow voltage changes, or "drift". With this system we are able to scan a range of 2·5 μ to 3·5 μ between 1 μ and 16 μ in a time of 14 sec. and to obtain at this scanning speed a resolving power which is sufficiently good for most work on complex molecules. Preliminary descriptions of an apparatus have already appeared (Daly and Sutherland, 1946).

§2. SPECTROMETER AND DETECTOR

A Nernst glower was used as a source of infra-red radiation, without special precautions other than shielding from draughts and the use of baretter lamps to drop the mains supply voltage. The radiation is collected by a mirror and focused on the spectrometer entrance slit through a rotating sectored disc, which gives a radiation intensity modulated in square wave form at about 20 c./s. The frequency is controlled by a mechanical governor and set by means of a stroboscope disc and neon lamp fed from the 50 c./s. mains. It may be varied over a range of 10 c./s. to 25 c./s., but it is found that certain sub-multiples of the mains frequency are to be avoided, as these may give rise to slow beats with residual mains hum in the amplifier. A frequency of exactly 20 c./s. has been found quite suitable for normal operation.

The spectrometer used was an old Bellingham and Stanley instrument having Littrow mounting of a 30° back-silvered rock-salt prism 5 cm. high and with a 3 cm. base. It has an aperture ratio of approximately F/12, which is rather low for this purpose, but was the only one readily available. The prism table drive has been modified (see figure 1) to operate from a continuously-rotating cam. This cam provides a uniform traverse of the prism table through an angle of about 1° for 300° rotation of the camshaft and a rapid flyback to the starting point over the remaining 60° of rotation. A screw control moves the cam roller relative to the prism table arm, thus giving a means of setting the start of the scan to any given wave-length. The angular traverse of the prism table during one scan may be varied by employing a selection of cams, but in our first model we have found that a cam cut to give a 300 cm.^{-1} scan in the 7 μ region is most generally useful.

The scan repetition frequency used is 3·5 per minute, which is well within the persistence limit of the cathode-ray-tube screen. The scanning speed is thus about 20 cm.^{-1}/sec. in the 7 μ region. For the present spectrometer and detector system, this appears to be the most satisfactory compromise. It is hoped that this may be raised to 100 cm.^{-1}/sec. with similar resolving power by the use of a spectrometer with larger aperture.

As mentioned above, the detector used is a thermistor bolometer developed by the Bell Telephone Laboratories. It consists of a flake of thermistor material with suitable current leads mounted on glass backing and having a receiver area of 2 mm. × 0·2 mm. The receiver operates in air but is sealed off from the atmosphere, radiation being admitted through a rocksalt window. This thermistor element is used in series with a compensating element matched to it in resistance/current characteristics, but not exposed to radiation. The two elements form the arms of a bridge (see figure 2) and from their junction the output signal is fed via a condenser to the grid of the first valve of the head amplifier.

If the radiation falling on the detector is modulated in this way, and the head amplifier responds to the modulation frequency but not to frequencies lower than a few cycles per second, the amplifier output will depend only on the amplitude of the *modulated* radiation and not on relatively slower radiation changes, or drift, caused by varying ambient temperature in any part of the system after the chopper. Sudden changes in radiation intensity, such as the switching on or off of lights, produce transient responses, which, however, are readily recognizable. In our system a modulation frequency of 20 cycles per second is used, a value which permits maximum signal-to-noise ratio to be reached, but which allows the amplifier pass-band to exclude interference from 50 c./s. mains hum.

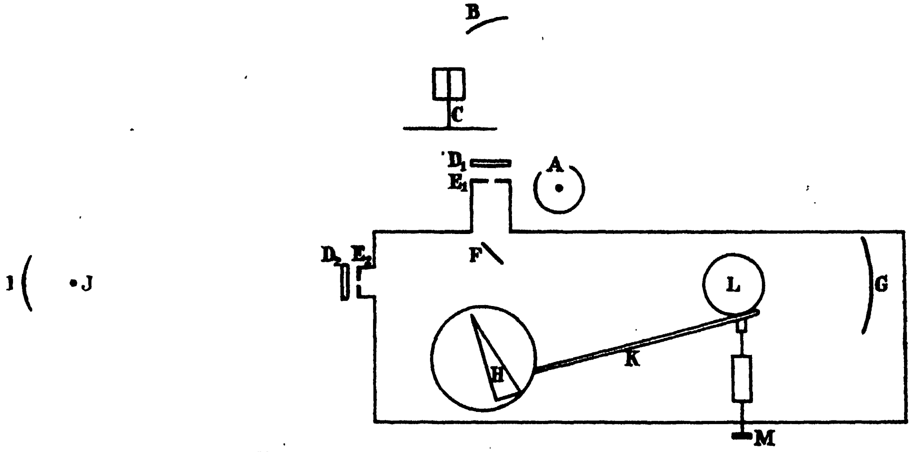


Figure 1. General lay-out of spectrometer.

A	Nernst filament.
B	Condensing mirror.
C	Chopping disk.
D_1. D_2	Alternative positions of absorbing material.
E_1	Entrance slit.
E_2	Exit slit.
F	Plane mirror.
G	Collimating mirror.
H	Prism.
I	Condensing mirror.
J	Bolometer.
K	Prism table arm.
L	Cam.
M	Control for cam setting.

The thermistor bolometer is operated with a d.c. bias voltage of between 100 and 150 across each element. Sensitivity to radiation increases with bias voltage, but so also does current noise in the thermistor flake. In addition, the flake resistance falls appreciably with increasing operating temperature, thus causing a fall in Johnson noise with increasing bias. A compromise must therefore be effected by choosing a bias voltage which is sufficiently high to give good sensitivity, but not so high that the current noise in the detector greatly exceeds the Johnson noise. It is found in practice that the curve of signal-to-noise ratio against bias voltage is fairly flat over the bias range 100–150 volts. Below this it falls off, and bias above this range tends to drive the bolometer into an unstable condition. Under optimum bias conditions, the resistance of the bolometer element is of the order of 2 megohms, so that, for noise purposes, the output signal is generated across a resistance of a megohm. The voltage sensitivity is approximately 250 μv. per μw. of radiation, interrupted at 20 c./s., for 100 volts D.C. bias per element.

§ 3. AMPLIFICATION AND DISPLAY

The present head amplifier (figure 2) was designed after some experience with that used in the original system (Daly and Sutherland, 1946). It has the following characteristics:—60 db. gain at 20 c./s., a response flat within 0·5 db. over the range 2 c./s. to 40 c./s., a noise level corresponding to a noise resistance of 100000 to 200000 ohms at the first grid, and an input impedance of not less than 5 megohms at 20 c./s.

Those features of the amplifier concerned with noise level will be discussed first. Thermal and current noise arising in the thermistor elements must be regarded as unavoidable. The remaining major sources of noise are (i) fluctuation in the power supplies to the bolometer bridge and to the amplifier; (ii) leakage through the coupling condenser between the thermistor bridge and the grid of the first amplifier valve, (iii) flicker noise in the early stages of the amplifier. Fluctuations in the power supplies may be reduced to a low value by adequate stabilizing and filtering circuits. The H.T. supplies are stabilized by the customary hard-valve series stabilizing circuit (see figure 2), the output from which is fed to a chain of gas-discharge valves which provides tappings at -150 0, $+150$ and $+300$ volts. The -150, $+150$ and $+300$ volt supplies are further smoothed by choke-condenser filters and by the decoupling networks in the head amplifier itself. The second source of noise was avoided by operating the thermistor bolometer bridge from the -150 and $+150$ volt lines. Arrangements are made for varying the bias while keeping the bridge mid-point approximately at earth potential. In this way the first coupling condenser (which is selected for low leakage) is not subjected to potential difference of more than a few volts.

Flicker noise in the valves is a well-known source of difficulty in designing low frequency amplifiers. The 6SJ7 valves used have relatively low flicker noise, and were operated at low screen voltages (about 10 v.) and with anode voltages of about 40 in the earlier stages to minimize ionization of any residual gas. It was not considered profitable to operate the valves at reduced heater voltage, as is often recommended, since this may shorten the useful life of the valves and was not found to give a significant improvement in signal-to-noise ratio. Carbon resistors carrying appreciable current were avoided where possible in the earlier stages of the amplifier. Hum from the 50 c./s. mains was reduced by employing a smoothed d.c. heater supply, and by complete electrostatic shielding of the bridge and first two amplifier stages.

Since, in the thermistor bridge, the signal voltage is generated across an effective impedance of a megohm, it is desirable that the input impedance of the amplifier should be as large as possible compared with this if loss of signal is to be avoided. This high input impedance was secured by using a cathode follower as the first stage, the cathode load being common to the first two valves. Since

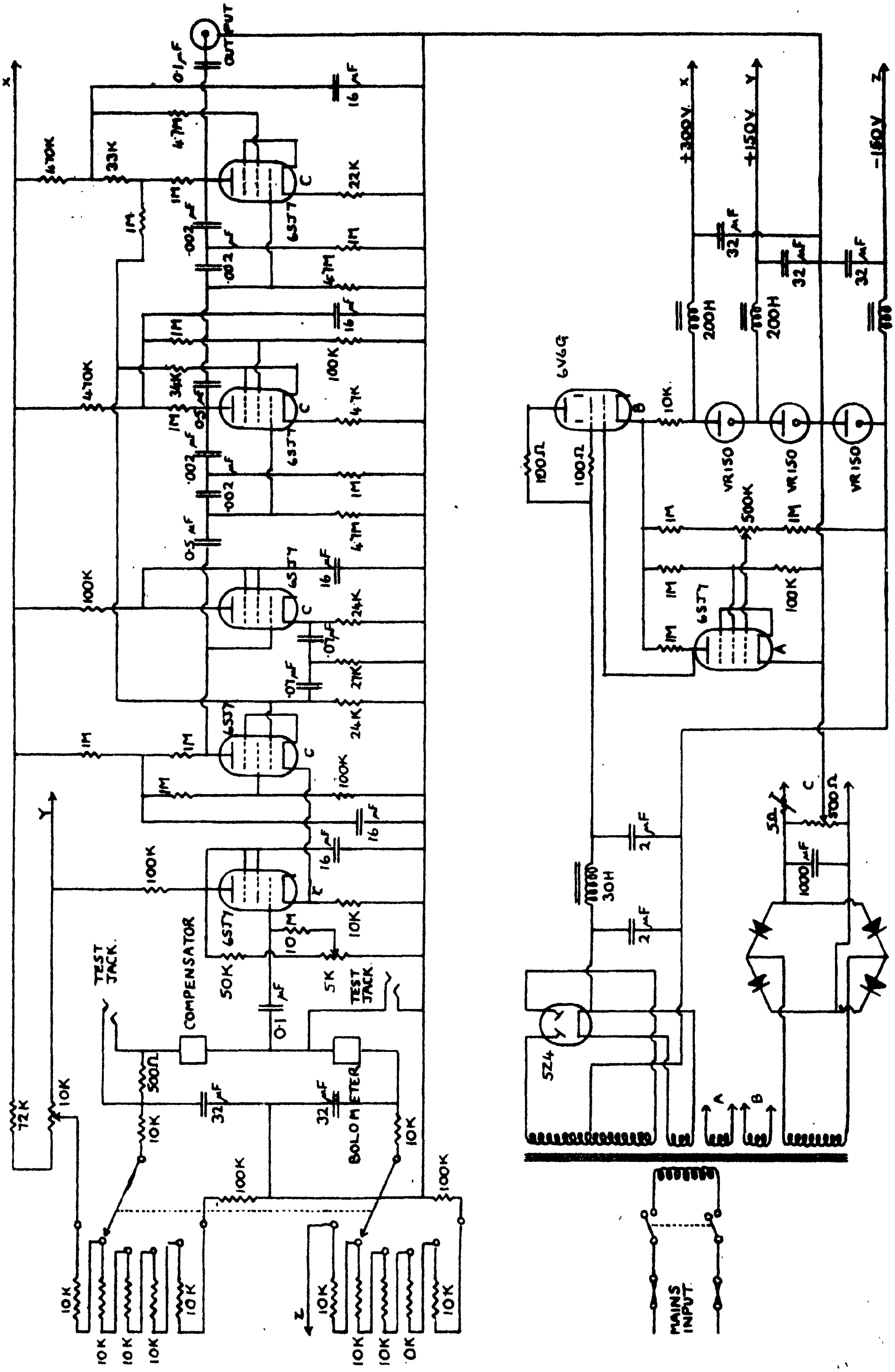


Figure 2. Head amplifier circuit.

negative feedback is applied to the control grid of the second stage, the cathode input impedance is relatively high, and the operation of the first valve is thus not impaired.

The response of the head amplifier was made wide in order that each energy pulse on the detector should give rise to a corresponding discrete voltage pulse at the amplifier output. Although in the circuits described below which follow the head amplifier, this band-width is greatly reduced by smoothing, it has been maintained at 40 c./s. up to detector level, as we intend to introduce shortly a system analogous to that of Baker and Robb (1943) for producing a direct plot of percentage transmission against wave-length instead of an energy trace. This system may involve the transmission of alternate pulses of widely different amplitudes, for the accurate reproduction of which a wide response band is essential.

Figure 3 shows the shape of the head amplifier response. This type of response was obtained by the use of negative feedback, via a high-pass T-section filter, between anode and control grid of each pentode amplifier stage (see figure 2) and by overall negative feedback from the anode of the final stage to the control grid of the first pentode stage, and from the cathode of the second pentode stage also to the control grid of the first. In spite of the multiplicity of feedback loops, the amplifier proved to be quite stable.

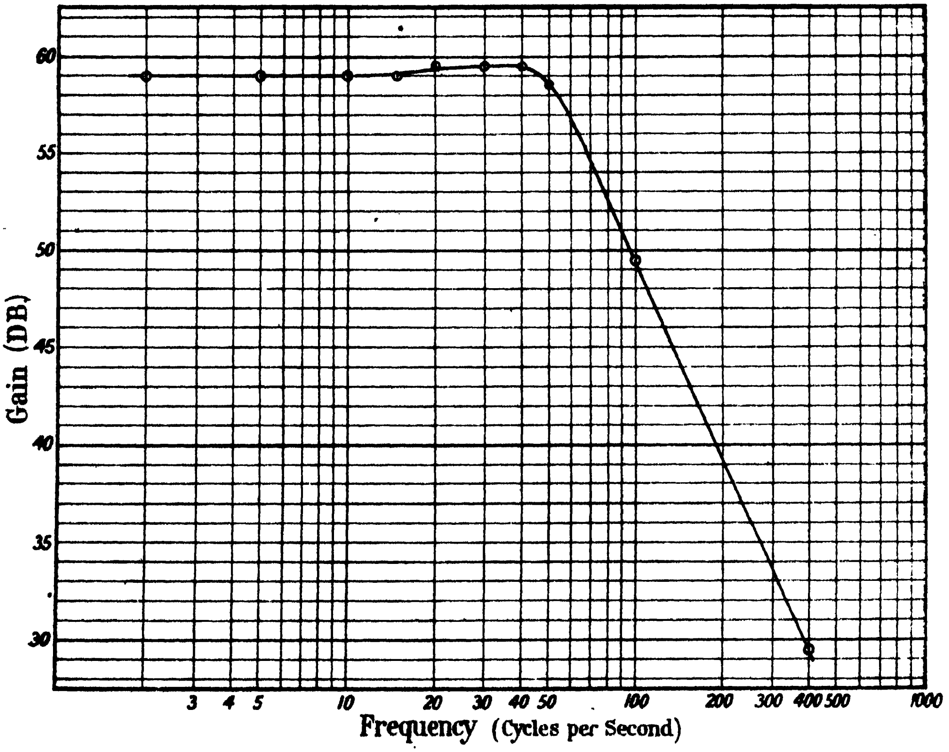


Figure 3. Head amplifier response.

The detector and head amplifier have been considered together as one unit—i.e. a radiation detector, whose output is at a sufficiently high level for it to be unaffected by the presence of normal noise in the succeeding electronic circuits. This unit has been arranged, as mentioned above, to give a close correspondence between input radiation pulses and output voltage pulses.

If following circuits are such that this correspondence is maintained up to the cathode-ray tube used for display, each radiation pulse will be represented by a proportional deflection (see plate 1 (*a*)). Alternatively, the radiation-chopping frequency may be treated as a carrier and detected, to yield finally a smoothed signal at a considerably narrower band-width (see plate 1 (*b*)). The former system is preferable where radiation energy is plentiful, and where it is desired to operate at high scanning speeds with low distortion of the form of the observed spectra. When energy is strictly limited, however, the signal-to-noise ratio at the display end may be increased by detecting and smoothing the interruption-frequency signal to a band-width of the order of 1/10 to 1/20 of that passed by the head amplifier. This considerable reduction in band-width is necessary if discrimination against ripple at the interruption frequency is to be achieved without a very elaborate filter system.

We have developed the second system, which is convenient for the qualitative examination of spectra, in the following manner:—Output signals from the head amplifier are fed through a gain control, having ten 6 db. steps, to an amplifier with 40 db. gain and a flat response from 2 c./s. to 40 c./s., which is similar in design to the head amplifier. The output from this second amplifier is detected by a diode and the output smoothed by means of a three-section, low pass, resistance-capacity filter. The 20 c./s. ripple is reduced in amplitude to 1 to 2% of the signal. The output of this filter, which has a band-width of about 2 c./s. is used to control the amplitude of a 500 c./s. square wave carrier (see figure 4). The circuit may be so adjusted that zero energy falling on the detector gives a very small carrier amplitude.

This 500 c./s. signal is then amplified by one stage with continuously variable gain (max. 35 db.) and finally by a paraphase amplifier in which the two anodes feed the Y plates of the cathode-ray tube. Between the variable gain stage and the grid of the paraphase amplifier, there is interposed either a d.c. restoring diode, or a clipping circuit with adjustable bias. The latter circuit is used if it is desired to have a sharp base-line which can be set precisely to zero energy on the detector, even in the presence of appreciable noise background. These circuits are shown in figure 4. An advantage of the 500 c./s. square wave carrier system is that the C.R.T. spot remains stationary for a considerable fraction of each cycle, so giving bright upper and lower limits to the pattern (see plate 1 (*b*)).

The cathode-ray tube is a 12-inch electrostatic deflection type, with a double screen of the type used in radar P.P.I. display units. A spot of the order of 0·5 mm. diameter is used and the maximum dimensions of the spectrum pattern are 15 cm. by 15 cm.

A time-base is provided by an accurate 100,000 ohm potentiometer linked to the spectrometer prism table drive. The potentiometer slider is connected to one X plate, and the other X plate to a voltage corresponding to the potentiometer mid-point. Continuous rotation of the potentiometer is possible, but for 60° the slider is not in contact with the winding. This 60° interval is used for the return of the prism table to its starting point, and during it the cathode-ray-tube trace is suppressed by a flasher and relay system driven from the same shaft as the time-base potentiometer and operating a brightness control.

§4. ILLUSTRATION OF SPECTRA

In plate 1 we give a comparison between the two methods of display. In the upper photograph we have the trace produced by the wide-band (40 c./s.) undistorted signal from the head amplifier. This is the same type of display as in our preliminary note in *Nature* (Daly and Sutherland, 1946). In the lower photograph we have the trace due to the narrow-band (2 c./s.) smoothed and re-modulated signal. The spectral region covered (roughly 1–4 μ) and the radiation intensity at the detector are the same for both plots. The source of radiation was a Nernst filament which has its maximum near 2 μ; superimposed atmospheric absorption bands at 1·4, 1·9 and 2·8 μ are clearly visible. In the upper photograph more detail is visible and this detail is reproducible with great accuracy, but from a psychological point of view the lower trace is preferable. While the paper was being written a note appeared (King *et al.*, 1946) giving an account of a copy of our original apparatus but with the introduction of smoothing. The authors of that note appear to consider that smoothing is always an advantage. They claim that it makes the detection of weak absorption bands easier. We would point out that this is not so, and *for the same scanning speed* smoothing must always involve some sacrifice of information. However,

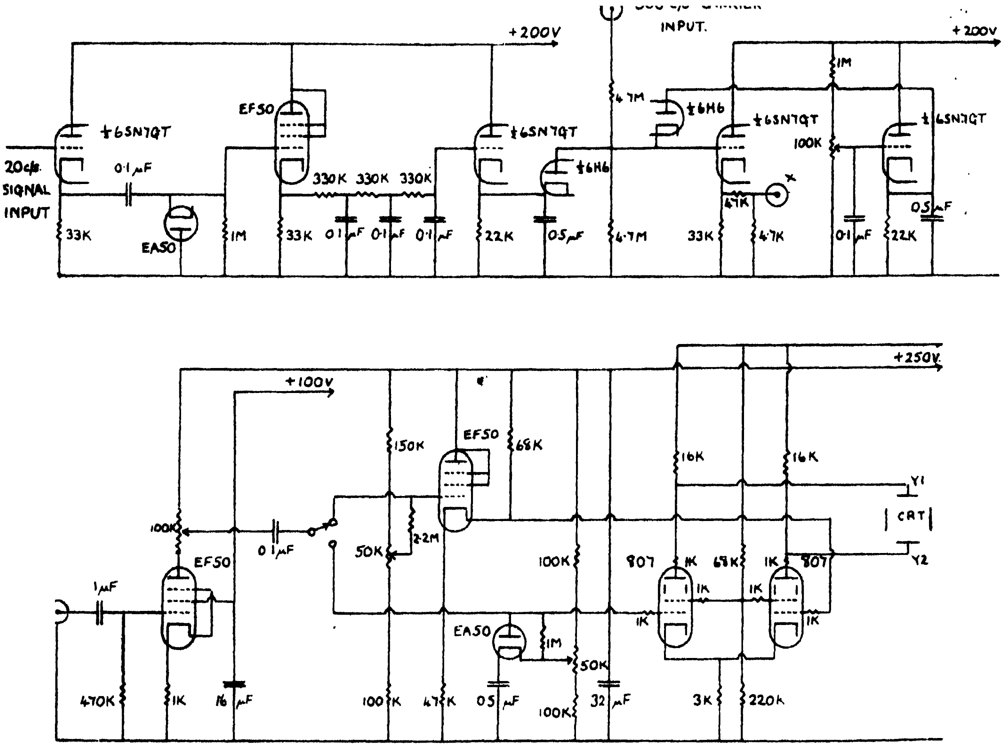
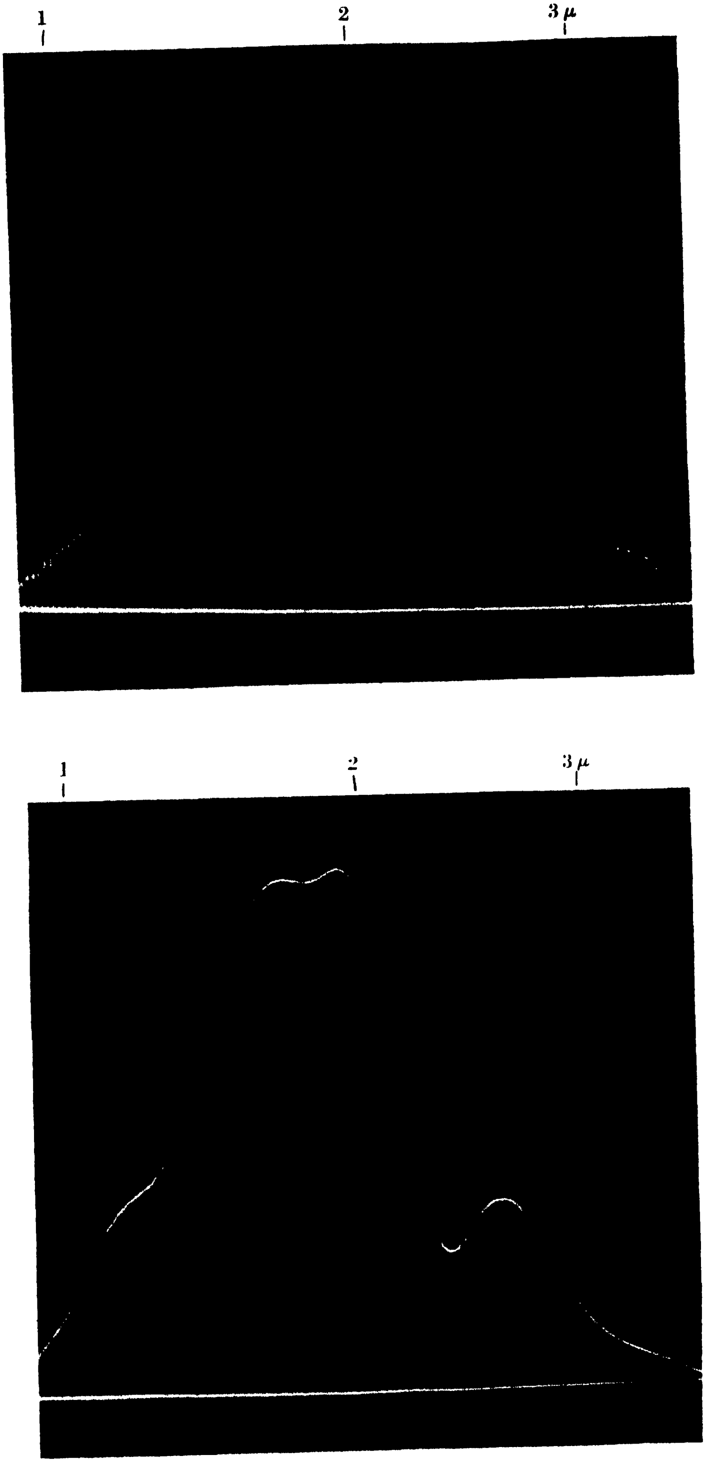
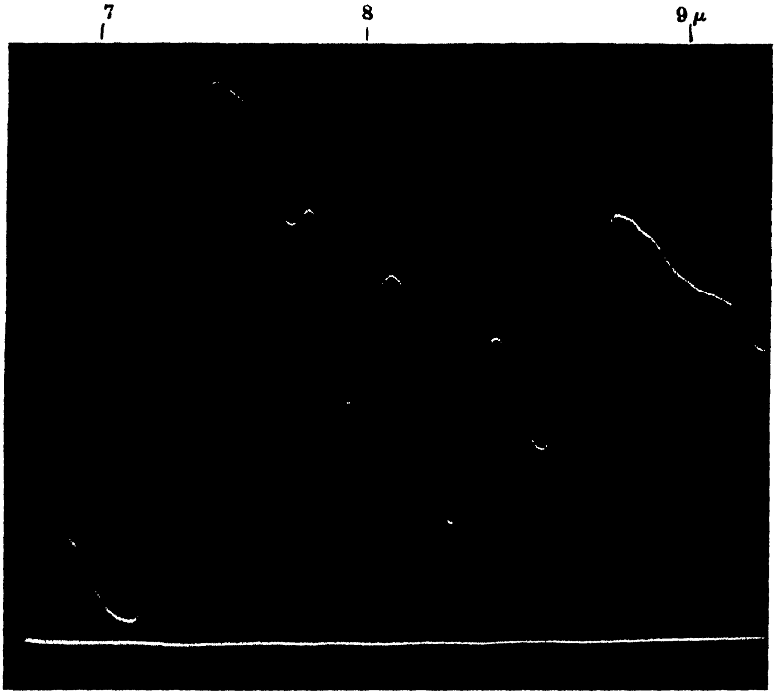


Figure 4. Circuits for smoothing, remodulation, and for final paraphase amplifier.

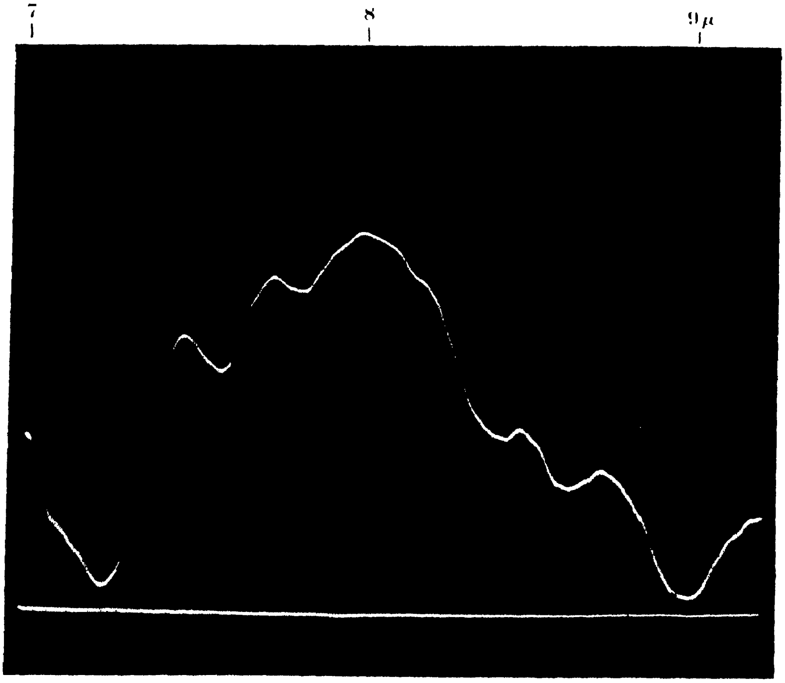


(b)

Plate 1. Emission spectrum from a Nernst filament.



(a) 2.2.4 Trimethylpentane



(b) 2.3.4 Trimethylpentane.

Plate 2. Absorption spectra of isomeric hydrocarbons.

for certain purposes this sacrifice may be unimportant compared to the psychological gain, and it is for this reason that we have developed both systems, which can be used interchangeably.

In plate 2 we illustrate the use of this instrument for rapid identification and analysis of chemicals. The upper photograph gives the absorption spectrum of 2.2.4 trimethylpentane over the approximate range 7–9 μ. The source of radiation was again a Nernst filament; the absorption minima in the energy curve are at 7·3, 7·8, 8·0, 8·3 and 8·5 μ and completely characterize this hydrocarbon. For comparison we give in the lower photograph the corresponding spectrum of the isomer 2.3.4 trimethylpentane with its equally characteristic series of bands at 7·3, 7·6, 7·9, 8·45, 8·6 and 8·9 μ. When one considers that these two plots were obtained in a time of approximately 30 seconds, and that there is every likelihood that this time can be reduced by a factor of 5 in the very near future, the potentialities of this instrument as a tool for chemical research and analysis need no further emphasis.

We have already used the spectroscope to follow dissociation phenomena and are at present studying certain chemical reactions. An account of this work will be more appropriately given at a later date in separate papers.

§ 5. FUTURE DEVELOPMENTS

In constructing this apparatus we have concentrated on simplicity and essentials. We considered the important thing was to show the practicability of such a spectroscope rather than to develop a very complicated instrument with elaborate refinements. Since an instrument of this type will have many diverse applications, various forms will have to be developed for special purposes. However, there are certain features to which we wish to draw attention, as we believe them to be of general interest to all future users.

Let us consider first the optical and mechanical parts of the spectroscope. Since in the infra-red the principal limitation to resolving power is determined by the energy passing through the exit slit, we must strive to maximize this energy for a given frequency range in the spectrum. This can be done by increasing the size of the prism, the aperture ratio and the length of the entrance slit.* Our present instrument employs a 30° prism with 3 cm. base with an effective aperture ratio F/12. We propose to use in future a 60° prism with 7 cm. base and an aperture ratio F/4·5. The limitations imposed here are solely the size of prism available and the cost of appropriate mirrors of the requisite focal length and optical accuracy. If the mirror is used off-axis, this last may be a serious consideration and it would be advisable to consider using an optical scheme such as that due to Pfund (1927) in which the collimating mirror is used on-axis.

In the present instrument we have used a single cam and a single scanning speed. The cam could be adjusted so that the 1° movement of the prism could take place in any part of the spectrum. This has the advantage of continuous variability, but is unsuitable in that it is designed really for the 6–8 μ region of

---

* We hope to publish a full analysis of these and other factors controlling the performance of this spectroscope in the very near future.

the spectrum. Thus, at shorter wave-lengths, the spectrum is too crowded, while at longer wave-lengths it is too spread out. An alternative method would be to have a series of, say, 5 cams to cover the range 1–16 μ with some overlap from one cam to the next. These cams could be interchangeable by means of a simple external control and calibrated to give a linear frequency scan within their respective ranges. Our original system has the great advantage that its flexibility allows more latitude in the comparison of absorption bands which might otherwise fall on different cams and we are retaining it with modifications which allow us to vary independently (1) the starting point of the scan, (2) the width of spectrum being scanned, and (3) the speed of scanning. At present no calibration marks are introduced in the spectrum displayed on the cathode-ray tube. A number of different methods of calibration are possible, but we propose to introduce one in which the trace is darkened momentarily as the spectrometer wave-length setting passes certain fixed values.

In the investigation of absorption spectra, the direct presentation of percentage transmission *v.* wave-length is desirable. The inclusion of this feature was decided upon at an early stage, and, as implied above, the system has been designed so that it may be added. The equipment now being set up for this purpose is in principle analogous to that of Baker and Robb (1943), but it will be seen that, in view of our higher interruption frequency and potentially higher scanning speed, it is desirable to substitute a purely electronic system for their electro-mechanical system of automatic gain control.

During the past six years a number of novel fast detectors have been developed for war purposes, in addition to the thermistor bolometer which we have used. Some of these, such as the superconducting bolometer (Andrews *et al.*, 1946), air cell detector (Weiss, 1946) and lead sulphide cell (Cashman, 1946; Starkiewicz, *et al.*, 1946), while not so convenient as the thermistor bolometer, offer considerable improvement in speed and sensitivity. Some of these are now under investigation by us.

Although the most striking feature of the present apparatus is that it gives a steady visible picture of the infra-red spectrum, its applicability to various problems will be restricted unless a record of the spectrum can be made, as desired, in tandem with its display. The obvious method of photography is rather clumsy and slow, so we are making arrangements to use a method of ultra-high-speed pen recording.

ACKNOWLEDGMENTS

This work forms part of a programme of fundamental research sponsored by the Hydrocarbon Research Group of the Institute of Petroleum. We wish also to express our thanks to the Ministry of Aircraft Production and to the Admiralty, the former for providing the thermistor bolometer and P.P.I. display tube, the latter for allowing one of us (E.F.D.) to work on this project while still in Admiralty employ (Oct.–Dec. 1945).

It is a pleasure to acknowledge the benefit of many stimulating discussions with Mr. Peter Fellgett, who also gave valuable assistance in the preliminary stages of this project.

REFERENCES

Andrews, D. H., Milton, R. M. and de Sorbo, W., 1946. *J. Opt. Soc. Amer.*, **36**, 353.
Baker, E. B. and Robb, C. D., 1943. *Rev. Sci. Instrum.*, **14**, 362.
Barnes, R. B., McDonald, R. S., Williams, V. Z. and Kinnaird, R. F., 1945. *J. Appl. Phys.*, **16**, 77.
Brattain, R. R. and Beeck, O., 1942. *J. Appl. Phys.*, **13**, 699.
Brattain, W. H. and Becker, J. A., 1946. *J. Opt. Soc. Amer.*, **36**, 354.
Cashman, R. J., 1946. *J. Opt. Soc. Amer.*, **36**, 356.
Daly, E. F. and Sutherland, G. B. B. M., 1946. *Nature, Lond.*, **157**, 547; *ibid.*, **158**, 591.
King, J., Temple, R. B. and Thompson, H. W., 1946. *Nature, Lond.*, **158**, 196.
McAllister, E. D., Matheson, G. L. and Sweeney, W. J., 1941. *Rev. Sci. Instrum.*, **12**, 314.
Pfund, A. H., 1927. *J. Opt. Soc. Amer.*, **14**, 337.
Starkiewicz, J., Sosnowski, L. and Simpson, O., 1946. *Nature, Lond.*, **158**, 28.
Sutherland, G. B. B. M. and Thompson, H. W., 1945. *Trans. Faraday Soc.*, **41**, 174.
Weiss, R. A., 1946. *J. Opt. Soc. Amer.*, **36**, 356.
Wright, Norman, 1941. *Ind. Eng. Chem.*, **13**, 1.

EQUIVALENT PATH AND ABSORPTION IN AN IONOSPHERIC REGION

By J. C. JAEGER,
The University of Tasmania

MS. received 24 *August* 1946

ABSTRACT. Numerical values for the equivalent path and absorption of a wireless wave incident vertically on a Chapman region are given for the cases of penetration and of reflection from either side of the region. The values are compared with those for the commonly assumed parabolic distribution of ionization. The application of the results to absorption measurements and to the reduction of experimental P', f curves is considered.

§ 1. INTRODUCTION

It is often convenient to be able to compare the relationship between equivalent path and frequency found experimentally, with the values calculated assuming a definite form for the distribution of ionization. For instance, Appleton (1937) has sometimes assumed that the ionization is parabolically distributed.

Chapman (1931) has made very general calculations of the distribution of ionization produced in the atmosphere by the incidence of monochromatic radiation from the sun. These are based on the assumption that the gaseous component responsible for absorbing the radiation is distributed exponentially with height, and allowance is made for the relative movement of the sun and the earth. It is shown here that the equivalent path and absorption of waves can be calculated for a region in which the ionization is distributed as indicated by Chapman's theory: such a region will be referred to below as a "Chapman region".

In studying absorption, Best and Ratcliffe (1928) have calculated values for a Chapman region, limiting the calculation to the case of waves of frequency much greater than the critical frequency for the region, so that it can be assumed that its refractive index is nearly unity. This limitation is not imposed here.

§2. DISTRIBUTION OF IONIZATION

The formulae relating ion production in the atmosphere to the mass absorption of solar radiation were first given by Lenard in 1911. In 1931 the complete theory of the region formation as a function of the position of the sun relative to the earth was given by Chapman (1931). He shows that if monochromatic radiation from the sun, of a wave-length that will ionize one of the atmospheric components, is incident upon the earth's atmosphere, the rate of ion production q at any height h above the earth's surface will be related to the zenith angle χ of the sun by the formula

$$q = q_0 \exp(1 - z - \sec\chi\, e^{-z}), \qquad \ldots\ldots(1)$$

where the earth is regarded as flat, and

$$z = \frac{h - h_0}{H}. \qquad \ldots\ldots(2)$$

Here z is the height above a datum level h_0 measured in terms of the scale height

$$H = \frac{kT}{mg}, \qquad \ldots\ldots(3)$$

where k is Boltzmann's constant, T is the absolute temperature, g is the acceleration of gravity, and m is the mean mass of the atmospheric molecules involved.

If it is assumed that the ions formed by the action of sunlight disappear by recombination at the rate αN^2 per c.c. per sec., where N is the number of ions existing per c.c. at any time, then N is determined during the day by the equation

$$\frac{dN}{dt} = q - \alpha N^2. \qquad \ldots\ldots(4)$$

It is known that dN/dt is small in the E-region in daylight; assuming as a first approximation that it vanishes, we have from (1) and (4)

$$N = \sqrt{(q/\alpha)} = N_0 \exp \tfrac{1}{2}(1 - z - \sec\chi e^{-z}), \qquad \ldots\ldots(5)$$

where $N_0 = \sqrt{(q_0/\alpha)}$.

The value of N given by (5) has a maximum

$$N_{\max} = N_0 \cos^{\frac{1}{2}}\chi \qquad \ldots\ldots(6)$$

at the level given by

$$e^{-z} \sec\chi = 1.$$

The value $\sec\chi = 1$ occurs at the equator at noon at the equinox. Thus N_0 is the maximum value of N at this place and time, and this value occurs when $z = 0$, that is, at a height h_0 above the earth's surface.

§3. ABSORPTION

When a radio wave is sent upwards at vertical incidence towards such an ionized region it is known from the theory of Lorentz that the refractive index μ and the absorption coefficient κ of the medium are given by

$$\mu^2 = 1 - \frac{N\epsilon^2}{\pi m f^2}, \qquad \ldots\ldots(7)$$

$$\kappa = \frac{\nu}{2c} \frac{1 - \mu^2}{\mu}, \qquad \ldots\ldots(8)$$

where f is the frequency of the wave, ϵ and m are the charge and mass of the ions,

c is the velocity of light, and ν is the collision frequency of the ions with neutral molecules. It is assumed that $(2\pi f)^2 \gg \nu^2$.

At vertical incidence the wave will be reflected from the region if there are sufficient ions at any level to reduce the value of μ to zero. This reflection occurs at the level at which $N = \pi m f^2/\epsilon^2$: if the effective ions are electrons, as is believed to be the case in E-region, this is equal to $1{\cdot}24 \times 10^4 f^2$, where f is the frequency in Mc./sec.

The value of f for which reflection occurs where $N = N_{\max}$ is called the critical frequency f_c for the region. It follows that

$$f_c^2 = \frac{\epsilon^2 N_{\max}}{\pi m} = \frac{\epsilon^2 N_0}{\pi m} \cos^{\frac{1}{2}} \chi. \qquad \ldots\ldots(9)$$

A small increase of f over f_c will cause the wave to penetrate the region.

The absorption coefficient κ involves the collision frequency ν of the electrons with neutral molecules. It is customary to assume that ν varies exponentially with height, so that

$$\nu = \nu_0 e^{-z}, \qquad \ldots\ldots(10)$$

where ν_0 is the value of ν at $z = 0$, that is at height h_0.

It is now desired to show how the total absorption varies with frequency and with the zenith angle of the sun for waves reflected from, or transmitted by, such a region.

The total absorption is $\int \kappa ds$, where the integral is taken along the whole trajectory of the wave. Thus from (2), (5), (7), (8), (10) we find for vertical incidence

$$\int \kappa ds = \int \frac{\nu}{2c} \frac{1-\mu^2}{\mu} ds$$

$$= \frac{\nu_0 f_0^2 H}{2c f^2} \int \frac{\exp \frac{1}{2}(1 - 3z - \sec \chi e^{-z})}{\sqrt{\{1 - (f_0^2/f^2) \exp \frac{1}{2}(1 - z - \sec \chi e^{-z})\}}} dz, \qquad \ldots\ldots(11)$$

where

$$f_0^2 = N_0 \epsilon^2 / \pi m. \qquad \ldots\ldots(12)$$

The integral can be put in a simple form involving the single parameter

$$\frac{f_0^2 \cos^{\frac{1}{2}} \chi}{f^2} = \frac{f_c^2}{f^2}. \qquad \ldots\ldots(13)$$

It is convenient also to use the notation

$$b = \frac{f_c^2}{f^2} \sqrt{(2e)}, \qquad \ldots\ldots(14)$$

and to make the substitution

$$y = (\tfrac{1}{2} \sec \chi)^{\frac{1}{2}} e^{-\frac{1}{2}z} \qquad \ldots\ldots(15)$$

in (11). This then becomes

$$\int \kappa ds = -\frac{\nu_0 H}{c \sec \chi} \int \frac{2 b y^2 e^{-y^2} dy}{\sqrt{(1 - b y e^{-y^2})}}. \qquad \ldots\ldots(16)$$

The integral has to be evaluated numerically between limits specified by the practical problem under consideration. There are three important cases to be considered.

(i) *A wave for which $f<f_c$ is propagated vertically upwards from a level at which μ is effectively unity*

In this case the wave is reflected from the level given by the greater root y_1 of the equation

$$bye^{-y^2}=1. \qquad \text{......(17)}$$

The value z_1 of z corresponding to this value of y is by (15)

$$z_1=\ln(\tfrac{1}{2}\sec\chi)-2\ln y_1, \qquad \text{......(18)}$$

and the height h_1 of the level of reflection above the earth's surface is by (2)

$$h_1=h_0+H\ln(\tfrac{1}{2}\sec\chi)-2H\ln y_1. \qquad \text{......(19)}$$

Values of y_1 for various values of f/f_c are given in table 1.

The total absorption in the wave thus reflected and returned to its starting-point is by (16)

$$\frac{\nu_0 H}{c\sec\chi}F_1\left(\frac{f}{f_c}\right), \qquad \text{......(20)}$$

where

$$F_1\left(\frac{f}{f_c}\right)=4b\int_{y_1}^{\infty}\frac{y^2e^{-y^2}dy}{\sqrt{(1-bye^{-y^2})}}. \qquad \text{......(21)}$$

Numerical values of the function F_1 for various values of f/f_c are given in table 1. Some remarks on the evaluation of the integral are made in § 7. The integral is divergent for $f=f_c$.

Table 1

	Reflection from below			Reflection from above			
f/f_c	y_1	$F_1(f/f_c)$	$P_1(f/f_c)$	y_2	$F_2(f/f_c)$	$P_2(f/f_c)$	$p_1(f/f_c)$
0·0	∞	4·000	0·0	0·0	0·0	5·545	0·0
0·05	2·805	4·252	0·354	0·001	0·000	5·545	0·003
0·1	2·526	4·310	0·438	0·004	0·000	5·545	0·010
0·15	2·344	4·360	0·510	0·010	0·000	5·546	0·023
0·2	2·204	4·408	0·578	0·017	0·001	5·547	0·041
0·25	2·087	4·454	0·646	0·027	0·003	5·550	0·064
0·3	1·985	4·502	0·716	0·039	0·006	5·555	0·094
0·35	1·893	4·552	0·789	0·053	0·012	5·564	0·129
0·4	1·809	4·606	0·867	0·069	0·020	5·577	0·172
0·45	1·730	4·663	0·952	0·088	0·033	5·597	0·222
0·5	1·654	4·726	1·046	0·108	0·051	5·625	0·281
0·55	1·581	4·796	1·150	0·132	0·076	5·664	0·351
0·6	1·510	4·875	1·269	0·158	0·110	5·717	0·432
0·65	1·440	4·966	1·407	0·188	0·156	5·790	0·528
0·7	1·369	5·073	1·570	0·221	0·219	5·890	0·643
0·75	1·297	5·203	1·771	0·258	0·305	6·027	0·782
0·8	1·222	5·367	2·027	0·300	0·426	6·220	0·958
0·85	1·142	5·588	2·376	0·350	0·604	6·507	1·189
0·9	1·053	5·913	2·901	0·411	0·887	6·971	1·522
0·92	1·013	6·099	3·207	0·441	1·057	7·252	1·708
0·94	0·969	6·346	3·616	0·475	1·287	7·639	1·950
0·95	0·944	6·506	3·885	0·494	1·438	7·895	2·105
0·96	0·918	6·704	4·222	0·516	1·629	8·221	2·296
0·98	0·854	7·341	5·328	0·570	2·249	9·304	2·901

(ii) *A wave for which $f > f_c$ is propagated vertically upwards from a level at which μ is effectively unity*

The wave will pass through the region and is assumed to be reflected downwards by a distinct upper region. The total absorption in the double passage of the lower region* is

$$\frac{\nu_0 H}{\mathbf{c}\sec\chi} F\left(\frac{f_c}{f}\right), \qquad \text{......(22)}$$

where

$$F\left(\frac{f_c}{f}\right) = 4b\int_0^\infty \frac{y^2 e^{-y^2}\,dy}{\sqrt{(1-bye^{-y^2})}}. \qquad \text{......(23)}$$

Numerical values of F are given in table 2.

Using (13) and (14), (22) may be put in the alternative form

$$\frac{\nu_0 H f_0{}^2}{\mathbf{c} f^2}(\sec\chi)^{-3/2}\,\Phi\left(\frac{f_c}{f}\right), \qquad \text{......(24)}$$

where

$$\Phi\left(\frac{f_c}{f}\right) = \frac{\sqrt{(2e)}}{b} F\left(\frac{f_c}{f}\right). \qquad \text{......(25)}$$

Values of the function Φ are also given in table 2. In the form (25) it appears

Table 2. Transmission

f_c/f	$F(f_c/f)$	$\Phi(f_c/f)$	$P(f_c/f)$	$p(f_c/f)$
0·0	0·000	4·133	0·000	0·000
0·05	0·010	4·136	0·010	0·003
0·1	0·041	4·146	0·042	0·013
0·15	0·094	4·164	0·094	0·030
0·2	0·168	4·188	0·169	0·055
0·25	0·264	4·221	0·267	0·087
0·3	0·384	4·262	0·390	0·127
0·35	0·528	4·313	0·540	0·177
0·4	0·700	4·374	0·720	0·236
0·45	0·900	4·447	0·934	0·308
0·5	1·134	4·534	1·187	0·394
0·55	1·403	4·639	1·485	0·497
0·6	1·715	4·765	1·839	0·621
0·65	2·078	4·917	2·261	0·771
0·7	2·501	5·105	2·771	0·956
0·75	3·005	5·341	3·398	1·189
0·8	3·616	5·649	4·195	1·493
0·85	4·387	6·073	5·256	1·911
0·9	5·436	6·711	6·796	2·543
0·92	5·992	7·080	7·658	2·909
0·94	6·689	7·570	8·780	3·396
0·95	7·120	7·889	9·496	3·713
0·96	7·637	8·287	10·375	4·108
0·98	9·183	9·562	13·125	5·378

---

* The absorption in the upper region is to be calculated from (20) and added to this.

that as $f_c/f \to 0$, $\Phi \to \sqrt{(2\pi e)} = 4 \cdot 133 \ldots$, and thus if $f \gg f_c$, the total absorption is

$$\frac{4 \cdot 13 \nu_0 H}{c} \frac{f_0^2}{f^2} (\sec \chi)^{-3/2}. \qquad \ldots\ldots(26)$$

This relation has been used by Best and Ratcliffe (1928) in a discussion of the experimental results. At lower frequencies it appears from (22) and (23) that the $(\sec \chi)^{-3/2}$ law in (26) does not hold, but it also follows quite generally that if f/f_c is a constant, that is, if a frequency is chosen which is a constant multiple of the critical frequency, the total absorption varies as $\cos \chi$. Thus to study the variation in absorption due to the variation in position of the sun it is desirable to experiment at a frequency which is a constant multiple of the critical frequency.

(iii) *A wave for which $f < f_c$ is propagated vertically downwards from above the layer from a level at which μ is effectively unity*

This case arises in an M reflection, or if waves are emitted from a source above the layer. Since the Chapman region is unsymmetrical, the absorption in this case will be different from that for a wave of the same frequency reflected from the lower side of the region, and it is of some interest* to determine the amount of this difference.

The level of reflection is now given by y_2, the least root of (17): values of this are given in table 1.

The total absorption is given by

$$\frac{\nu_0 H}{c \sec \chi} F_2\left(\frac{f}{f_c}\right), \qquad \ldots\ldots(27)$$

where

$$F_2\left(\frac{f}{f_c}\right) = 4b \int_0^{y_2} \frac{y^2 e^{-y^2}\,dy}{\sqrt{(1 - bye^{-y^2})}}. \qquad \ldots\ldots(28)$$

Values of F_2 are given in table 1. It appears that the absorption is much less than that for a wave of the same frequency reflected from the lower side of the region.

§4. THE EQUIVALENT PATH

The equivalent path of the waves in the region is

$$\int \frac{ds}{\mu}, \qquad \ldots\ldots(29)$$

where μ is given by (5) and (7). Since, in all the cases in which we are interested the path begins and ends in regions where μ is nearly unity, it is better to evaluate the excess of equivalent path over distance traversed, that is

$$\int \left(\frac{1}{\mu} - 1\right) ds. \qquad \ldots\ldots(30)$$

---

* It has been shown by Bagge (1943) that the ion density in the upper parts of the region will be reduced by diffusion to values much lower than the theoretical Chapman values. Thus the results given here will not correspond to the practical case at low frequencies.

Using the notation of § 3, and making the substitutions used there, this becomes

$$-2H\int \frac{dy}{y}\left\{\frac{1}{\sqrt{(1-bye^{-y^2})}}-1\right\} \qquad \ldots\ldots(31)$$

Then in the three cases of § 3 the results are as follows:

Case (i). $f<f_c$. *Reflection from below*

The excess of equivalent path over distance traversed with which the wave returns to its starting-point is $HP_1(f/f_c)$, where

$$P_1\left(\frac{f}{f_c}\right)=4\int_{y_1}^{\infty}\frac{dy}{y}\left\{\frac{1}{\sqrt{(1-bye^{-y^2})}}-1\right\} \qquad \ldots\ldots(32)$$

and y_1 is the greater root of (17). y_1 and $P_1(f/f_c)$ are given in table 1. The actual equivalent path from the earth's surface to the layer and back is, by (19),

$$2h_0+2H\ln\left(\tfrac{1}{2}\sec\chi\right)-4H\ln y_1+HP_1(f/f_c). \qquad \ldots\ldots(33)$$

Case (ii). $f>f_c$. *Penetration*

The excess of equivalent path over distance in a double passage of the region is $HP(f_c/f)$, where

$$P\left(\frac{f_c}{f}\right)=4\int_{0}^{\infty}\frac{dy}{y}\left\{\frac{1}{\sqrt{(1-bye^{-y^2})}}-1\right\}. \qquad \ldots\ldots(34)$$

Values of $P(f_c/f)$ are given in table 2.

Case (iii). $f<f_c$. *Reflection from above*

The excess of equivalent path over distance is $HP_2(f/f_c)$, where

$$P_2\left(\frac{f}{f_c}\right)=4\int_{0}^{y_2}\frac{dy}{y}\left\{\frac{1}{\sqrt{(1-bye^{-y^2})}}-1\right\}, \qquad \ldots\ldots(35)$$

and y_2 is the least root of (17). Values of y_2 and P_2 are given in table 1; it appears that P_2 is much greater than P_1.

§ 5. COMPARISON WITH RESULTS FOR A PARABOLIC DISTRIBUTION OF IONIZATION OVER A REGION OF THICKNESS $2y_m$

The parabolic distribution has been widely studied because of its simplicity. The results have been given by Appleton (1937) and are as follows.

In the case $f>f_c$ in which the wave penetrates the region, the excess of equivalent path over distance in a double passage is $y_m p(f_c/f)$, where

$$p(f_c/f)=\frac{2f}{f_c}\ln\frac{f+f_c}{f-f_c}-4. \qquad \ldots\ldots(36)$$

The quantity $p(f_c/f)$ is shown in the last column of table 2 for comparison with the corresponding quantity $P(f_c/f)$ for the Chapman region. Since equivalent paths in the two cases are expressed in terms of the thickness $2y_m$ and the scale height H of their respective regions, direct comparison of the results is not possible, but the ratio $p:P$ may be seen to decrease steadily from 0·41 when $f_c/f=0{\cdot}98$ to 0·32 when $f_c/f=0{\cdot}1$.

It may be remarked that the Chapman distribution (5) has two points of inflexion, and that the vertical distance between these is independent of χ and equal to

$$H \ln \frac{2+\sqrt{3}}{2-\sqrt{3}} = 2{\cdot}63H.$$

Thus if a Chapman region were approximated to by a parabolic distribution, the thickness $2y_m$ of the latter would be of the order of, but greater than, $2{\cdot}63H$.

For the parabolic distribution, if $f<f_c$, the wave is reflected at a height

$$y_m \frac{f_c-\sqrt{(f_c^2-f^2)}}{f_c} \qquad \ldots\ldots(37)$$

above the bottom of the layer, and the equivalent path from the bottom of the layer to the level of reflection and back is

$$\frac{2y_m f}{f_c} \ln \frac{f_c+f}{f_c-f}. \qquad \ldots\ldots(38)$$

From (37) and (38) the excess of equivalent path over distance for a wave reflected from the layer may be calculated. Writing this in the form $y_m p_1(f/f_c)$, the quantity $p_1(f/f_c)$ is given in the last column of table 1 for comparison with the corresponding quantities $P_1(f/f_c)$ and $P_2(f/f_c)$ for reflection from the under and upper sides of a Chapman region. It appears that there are large differences between the three curves.

§6. COMPARISON WITH EXPERIMENTAL P', f CURVES

There are two cases to be considered, according to whether the wave is reflected from the lowest layer of the ionosphere or passes through this and is reflected from a higher layer.

(i) *Reflection from a region of scale height H at height h_0*

In this case the equivalent path P' is by (33)

$$P' = 2h_0 + 2H \ln(\tfrac{1}{2}\sec\chi) - 4H \ln y_1 + HP_1(f/f_c), \qquad \ldots\ldots(39)$$

where y_1 and P_1 are given in table 1. Measurements of an experimental P', f curve at two values of f/f_c give two equations from which h_0 and H can be found.

(ii) *The wave passes through a region of scale height H_E and critical frequency f_E, and is reflected from a higher region of scale height H at height h_0 whose critical frequency is f_c*

In this case

$$P' = 2h_0 + 2H \ln(\tfrac{1}{2}\sec\chi) - 4H \ln y_1 + HP_1(f/f_c) + H_E P(f_E/f), \qquad \ldots\ldots(40)$$

where y_1 and P_1 are given in table 1 and P in table 2. In this case measurement of three points on an experimental P', f curve will give the values of h_0, H and H_E.

§7. EVALUATION OF THE INTEGRALS

This offers no particular difficulty. The integrals are readily transformed into forms suited to numerical integration, and series expansions are also available which give satisfactory convergence for most of the values of f/f_c.

In the case of penetration the series expansions are

$$F(f_c/f) = \sum_{n=1}^{\infty} c_{n-1} b^n \Gamma(\tfrac{1}{2}n) n^{-\frac{1}{2}n}, \qquad \ldots\ldots(41)$$

$$P(f_c/f) = 2 \sum_{n=1}^{\infty} c_n b^n \Gamma(\tfrac{1}{2}n) n^{-\frac{1}{2}n}, \qquad \ldots\ldots(42)$$

where

$$c_n = \frac{1\cdot 3 \ldots . (2n-1)}{2\cdot 4 \ldots . (2n)}, \quad c_0 = 1. \qquad \ldots\ldots(43)$$

In the reflection case the roots y_1 and y_2 of (17) may be found from tables of ye^{-y^2} or of the derivative of the error function. The series expansions are rather more complicated; for example, for reflection from the lower side of the region,

$$P_1(f/f_c) = 2 \sum_{n=0}^{\infty} \frac{\Gamma(n+\frac{1}{2}) c_{2n+1} b^{2n+1}}{(2n+1)^{(2n+1)/2}} \operatorname{erfc}[y_1(2n+1)^{\frac{1}{2}}]$$
$$+ 4 \sum_{n=1}^{\infty} \frac{1}{(2y_1^2)^n} \sum_{r=2n}^{\infty} \frac{(r;\, n) c_r}{r^n}, \qquad \ldots\ldots(44)$$

where

$$(r;\, n) = (r-2)(r-4) \ldots . (r-2n+2), \quad (r;\, 1) = 1,$$

and c_n is defined in (43). The most important of the numerical series which appear in (44) as coefficients of y_1^{-2n} can be summed accurately; the others are easily estimated numerically.

For reflection from the upper side of the region a series for y_2 is available (Bromwich (1926), p. 161), and this leads easily to series expansions for F_2 and P_2.

§8. EXTENSION TO HIGHER VALUES OF THE COLLISION FREQUENCY

Throughout the above work it has been assumed that $\nu^2 \ll 4\pi^2 f^2$. This relation is normally satisfied in E- and F-regions, but it is probable that it will not hold in D-region. This point is of some importance in connection with the study of absorption. Most workers, e.g. White and Straker (1939), use the Best and Ratcliffe formula (26) for the study of diurnal and seasonal variation of absorption, remarking that if much of the absorption takes place in D-region this formula, and in particular the $(\cos\chi)^{3/2}$ law, will not hold good.

It is thus of some interest to study absorption at relatively high values of the collision frequency. There is no difficulty in extending the above numerical work to this case, but for a region in which μ is nearly unity, a simple extension of the Best and Ratcliffe formula (26) is easily obtained, and this should be sufficient for the discussion of experimental results.

To derive this we assume $f \gg f_c$, so that μ is nearly unity, but we do not neglect $\nu^2/4\pi^2 f^2$. Then (8) is replaced by

$$\kappa = \frac{2\pi\epsilon^2 N\nu}{mc(4\pi^2 f^2 + \nu^2)}.$$

Using the values (5) and (10) of N and ν, this gives for a double traverse of the region

$$\int \kappa\, ds = \frac{4\pi\epsilon^2 N_0 \nu_0 H}{mc} \int_{-\infty}^{\infty} \frac{\exp(\frac{1}{2} - \frac{3}{2}z - \frac{1}{2}\sec\chi e^{-z})\, dz}{4\pi^2 f^2 + \nu_0^2 e^{-2z}}.$$

Making the substitution (15), this becomes

$$\int \kappa ds = \frac{H\nu_0 f_0^2}{cf^2}(\sec\chi)^{-3/2} I, \qquad \ldots\ldots(45)$$

where

$$I = 4\beta^2\sqrt{(2e)}\int_0^\infty \frac{y^2 e^{-y^2}\,dy}{y^4+\beta^2}, \qquad \ldots\ldots(46)$$

and

$$\beta = \frac{\pi f \sec\chi}{\nu_0}. \qquad \ldots\ldots(47)$$

It is found after some reduction that

$$I = 4\pi\beta^{\frac{3}{2}}e^{\frac{1}{2}}\{[\tfrac{1}{2}-C(\beta)]\cos\beta + [\tfrac{1}{2}-S(\beta)]\sin\beta\}, \qquad \ldots\ldots(48)$$

where $C(\beta)$ and $S(\beta)$ are the Fresnel integrals

$$C(\beta) = \frac{1}{\sqrt{(2\pi)}}\int_0^\beta \frac{\cos t}{\sqrt{t}}\,dt, \quad S(\beta) = \frac{1}{\sqrt{(2\pi)}}\int_0^\beta \frac{\sin t}{\sqrt{t}}\,dt, \qquad \ldots\ldots(49)$$

which are tabulated in Jahnke-Emde (1933).

For large values of β, that is $2\pi f \gg \nu_0$, I has the value 4·133, and we regain (26). The behaviour of I for other values of β may be seen from table 3.

Table 3

β	I	β	I	β	I
0·0	0·000	2·5	3·093	8	3·931
0·5	1·109	3·0	3·287	10	3·996
1·0	1·919	4·0	3·547	15	4·067
1·5	2·463	5·0	3·706	20	4·095
2·0	2·828	6·0	3·810	∞	4·133

Since I is an increasing function of β and thus of $\sec\chi$, it follows that if $\int \kappa ds$ is plotted against $(\cos\chi)^{3/2}$ the resulting curve will not be the straight line given by (26) but will be concave downwards, the amount of the concavity being an indication of the amount of absorption taking place in regions where the collision frequency is high. The presence of a D-region of Chapman type will also have the effect of diminishing the ratio of summer to winter absorption.

Finally it should be remarked that although the effects of the earth's magnetic field have been neglected throughout, the results apply in the usual way to the important case of quasi-longitudinal propagation provided f is replaced by $f \pm f_L$.

ACKNOWLEDGMENT

I am indebted to Dr. F. W. G. White for introducing me to the subject, and for much helpful discussion; this work was begun in collaboration with him in the Division of Radiophysics of the Council for Scientific and Industrial Research, Australia.

REFERENCES

Appleton, E. V., 1937. *Proc. Roy. Soc.*, A, **162**, 451.
Bagge, E., 1943. *Phys. Z.*, **44**, 163.
Best, J. E. and Ratcliffe, J. A., 1928. *Proc. Phys. Soc.*, **50**, 233.
Bromwich, T. J. I'A., 1926. *Theory of Infinite Series*, Ed. 2 (Macmillan).
Chapman, S., 1931. *Proc. Roy. Soc.*, A, **132**, 353; 1931. *Proc. Phys. Soc.*, **43**, 26.
Jahnke-Emde, 1933. *Tables of Functions*, Ed. 2 (Teubner).
White, F. W. G. and Straker, T. W., 1939. *Proc. Phys. Soc.*, **51**, 865.

OBLIQUE RADIO TRANSMISSION IN THE IONOSPHERE, AND THE LORENTZ POLARISATION TERM

By W. J. G. BEYNON,*

Communication from the National Physical Laboratory

* Now at University College, Swansea

MS. received 19 *September* 1946 †

ABSTRACT. The application of the Sellmeyer and Lorentz dispersion formulae to the problem of oblique reflection of radio waves from a parabolic type of layer is discussed. The work of Ratcliffe on the problem is extended and shown to be consistent with that of N. Smith. This theoretical analysis has been applied to a large number of experimental measurements of the maximum usable frequency at oblique incidence.

The results of two sets of measurements of the maximum usable frequency over distances of 1000 km. and 700 km. are in close agreement with values calculated from the Sellmeyer type of formula. In the case of the 1000 km. observations, the divergence between calculated and observed values is less than 0·2 of what it should be if the Lorentz formula applied. In the case of more accurate measurements over a distance of 700 km., the mean divergence between calculated and observed values is only about 0·03 of that which should occur on the basis of the Lorentz formula. These experimental results are thus in agreement with the theoretical conclusion of Darwin, that the Sellmeyer type of formula is applicable to the case of refraction of radio waves in the ionosphere.

§ 1. INTRODUCTION

The question whether the Sellmeyer formula

$$\mu^2 - 1 = \frac{-4\pi Ne^2}{mp^2}$$

or the Lorentz formula

$$\frac{3(\mu^2 - 1)}{\mu^2 + 2} = \frac{-4\pi Ne^2}{mp^2}$$

should be applied to the case of refraction in an ionised medium has been examined from the theoretical point of view by Darwin (1934, 1943). From the experimental side the problem has been considered by Booker (1938), Martyn (1938, 1939), Ratcliffe (1939), Smith (1941) and others. Darwin's theoretical analysis indicates that the simple Sellmeyer formula should be applicable, but the available experimental evidence has not yet been conclusive. From experiments on medium frequencies, Booker and Berkner concluded that it would be extremely difficult to interpret their results in terms of the Sellmeyer theory. Martyn and Munro also considered frequencies near the gyro-magnetic frequency, but concluded that the Sellmeyer theory was applicable. The experimental and theoretical work of Ratcliffe led to no conclusive result, whilst similar work by Smith again suggests that the Sellmeyer formula was appropriate to ionosphere theory.

In the present paper the theoretical work of Ratcliffe has been extended and shown to be consistent with that of Smith. The results of some oblique-incidence

---

† This paper was originally communicated to the Radio Research Board in 1943.

ionospheric measurements are also considered. These results appear capable of explanation only in terms of the Sellmeyer formula.

§2. EARLIER WORK OF RATCLIFFE AND OF SMITH

The effect of including or omitting the so-called Lorentz polarisation term in oblique incidence problems has been discussed by Ratcliffe (1939) and by Smith (1941). Ratcliffe considered an ionosphere with a parabolic gradient of ionisation and concluded that at very short or very long distances there would be no marked difference between the magnitude of the maximum usable frequency calculated from either formula. He states that for region F_2 the greatest divergence between the two cases could only amount to about 2%. This would occur for a sender-receiver distance of about 500 km., so that in practice the accuracy of measurement would probably not be sufficient to decide between the two theories. In fact, an attempt to apply the analysis to the results of some oblique-incidence experiments for a sender-receiver distance of 464 km. was unsuccessful. On the other hand, Smith (1941), from an analysis of the same problem, concluded that the divergence between the two calculated values of the maximum usable frequency would progressively increase with increasing distance. For limiting distances of 1500 km. to 3000 km., Smith's analysis indicated that for transmissions by region E there should be a difference between maximum usable frequencies calculated from the Sellmeyer and Lorentz formulae of some 19% and for transmission by region F_2 a difference of about 17%. From the calculations given below, it seems that the different results given by Ratcliffe and by Smith arose from the fact that Ratcliffe confined his calculations to comparatively short distances. Applied to larger distances the formulae of Ratcliffe show that the difference between maximum usable frequencies calculated from the Sellmeyer and Lorentz formulae is quite large and of the order stated by Smith.

§3. CALCULATION OF THE MAXIMUM USABLE FREQUENCY (M.U.F.) FACTOR CURVE

(a) *Case of plane earth and ionosphere. Lorentz formula*

It will be convenient to reconsider briefly the analysis given by Ratcliffe and to examine the extension of his formulae to greater distances. The effect of the earth's magnetic field will be neglected, and initially the earth and ionosphere are considered to be plane. The nomenclature will be similar to that used by Ratcliffe. The ionosphere is assumed to have a parabolic gradient of ionisation density, and the scale of height is arranged so that the semi-thickness is unity. The Lorentz formula can then be written

$$\mu^2 = 2\left(\frac{z^2 - a^2}{b^2 - z^2}\right),$$

where μ is the refractive index at distance z below the level of maximum ionisation density;

$$a^2 = 1 - \frac{f^2}{f^{02}}, \qquad b^2 = 1 + \frac{2f^2}{f^{02}};$$

f is the frequency under consideration and f^0 the normal-incidence critical frequency of the reflecting region.

The horizontal range D_1 corresponding to the portion of the path in the layer of a frequency f incident at angle i_0 (figure 1) is then given by

$$D_1=2\int_{\mu=1}^{\mu=s}\tan i\,.\,dz=2\int_{z=1}^{z=a}\frac{\sin i_0\,.\,dz}{\sqrt{\mu^2-s^2}}$$

$$=2s(2+s^2)^{-\frac{1}{2}}(1+2\epsilon^2)^{\frac{1}{2}}\Big[F(k,\chi)-E(k,\chi)\Big]_{\alpha}^{\pi/2}, \qquad \ldots\ldots(1)$$

where $s=\sin i_0, \qquad \epsilon=f/f^0,$

$$\alpha=\sin^{-1}\sqrt{\frac{b^2-1}{b^2-c^2}}=\sin^{-1}\sqrt{\frac{2+s^2}{3}},$$

$$k^2=\frac{6\epsilon^2}{(2+s^2)(1+2\epsilon^2)}, \quad c^2=\frac{2a^2+s^2b^2}{2+s^2}.$$

$F(k,\chi)$ and $E(k,\chi)$ can readily be evaluated from the usual tables of elliptic integrals.

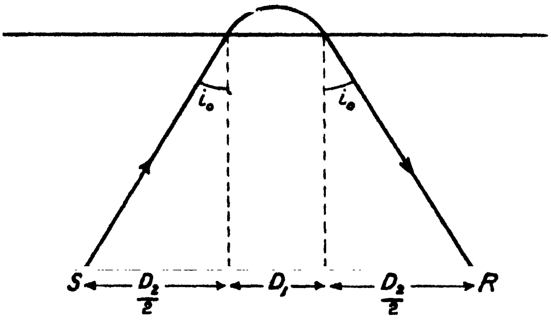


Figure 1.

The total range D is given by

$$D=D_1+2h_0\tan i_0, \qquad \ldots\ldots(2)$$

h_0 being the height of the lower edge of the layer above the ground. We now plot the relation between D and i_0 for a particular fixed value of ϵ and so find when the range is a minimum. This gives us the maximum usable frequency for that particular range.

(b) *Case of plane earth and ionosphere. Sellmeyer formula*

For a parabolic layer in which the Lorentz term is assumed to be zero, the horizontal range is given by

$$D=D_1+2h_0\tan i_0=\epsilon\sin i_0\log_e\frac{1+\epsilon\cos i_0}{1-\epsilon\cos i_0}+2h_0\tan i_0. \qquad \ldots\ldots(3)$$

In this case the M.U.F. curve can readily be deduced without plotting (D, i_0) curves by the method given by Appleton and Beynon (1940, 1947).

(c) *The equivalent height measured at normal incidence*

In practice the calculation of the M.U.F. factor curve is based on normal-incidence equivalent-height measurements. It is therefore necessary to examine the form of the (equivalent-height, frequency) curve to be expected from each of the two theories.

Lorentz formula

In this case Ratcliffe has shown that for a layer of parabolic type the equivalent height in the layer at normal incidence is given by

$$h'(\epsilon, 0) = \sqrt{\frac{2}{1+2\epsilon^2}}\Big[(1+4\epsilon^2)E(k,\theta) - (1-2\epsilon^2)F(k,\theta)\Big]_{\theta=0\cdot956}^{\theta=\pi/2} - \tfrac{1}{3}k^2, \qquad \ldots\ldots(4)$$

k and ϵ having the same significance as before.

Sellmeyer formula

In this case the corresponding formula is

$$h'(\epsilon, 0) = \frac{\epsilon}{2}\log_e\left(\frac{1+\epsilon}{1-\epsilon}\right). \qquad \ldots\ldots(5)$$

Figure 2 shows the equivalent height in the layer calculated from equations (4) and (5). It will be noted that the difference between the two cases is small. If we denote the semi-thickness of the region by y_m then the maximum difference at any frequency only amounts to about $0{\cdot}11y_m$. In fact the change from the Sellmeyer curve to the Lorentz curve can be simulated by small changes in the level of maximum ionisation density and in the semi-thickness of the region. Examination of figure 2 shows that if application of the Lorentz formula to any given experimental vertical incidence (h', f) curve, yields a level of maximum ionisation, say $h_0 + y_m$ and a semi-thickness y_m, then application of the Sellmeyer formula to the same curve would approximately yield a level of maximum ionisation of $h_0 + 0{\cdot}89y_m$

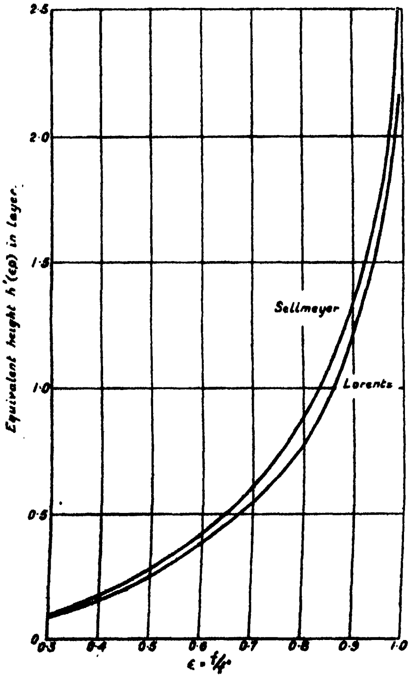


Figure 2. Normal incidence (h', ϵ) curves. Lorentz and Sellmeyer formulae.

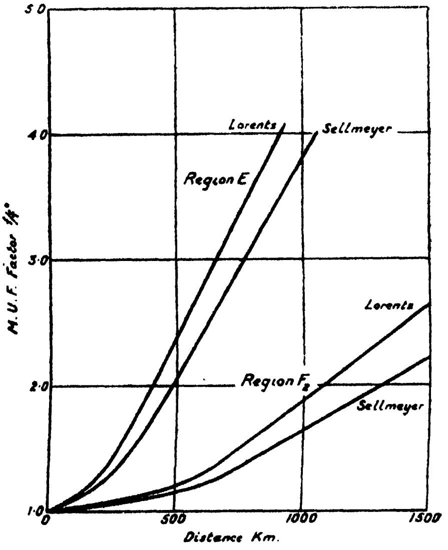


Figure 3. M.U.F. factor curves. Plane earth and ionosphere.

and a semi-thickness of $0{\cdot}93y_m$. Hence, in comparing M.U.F. factor curves calculated from the two theories, we must really compare the Lorentz M.U.F. factor curve for a level of maximum h_0+y_m and semi-thickness y_m with the Sellmeyer curve corresponding to a level of maximum $h_0+0{\cdot}89y_m$ and semi-thickness $0{\cdot}93y_m$. Figure 3 shows four M.U.F. factor curves, two for region E and two for typical region F_2 conditions. One curve in each pair corresponds to the Lorentz formula and one to the Sellmeyer formula. It will be noted that for distances less than about 500 km. for region F_2, the difference between the two cases is small, but the divergence steadily increases with distance, and at 1000 km. amounts to 12%. For region E the divergence is much larger than for region F at all distances. These calculations have assumed a plane earth and ionosphere, and rigid application of the results is really confined to distances for which earth-curvature effects are negligible. Any correction to these results for earth and ionosphere curvature will, however, affect the Sellmeyer and Lorentz curves to about the same order, and it is found that the ratios of M.U.F. factors to be deduced from the curves in figure 3 are really valid up to about 1000 km. We now consider some modifications of these formulae which will compensate approximately for the effect of earth and ionosphere curvature.

(d) *Case of curved earth and ionosphere. Lorentz formula*

Again we consider a parabolic type of layer, the height of the lower edge being h_0.

Let μ represent the refractive index at radial distance r from the centre of the earth. Then if χ represents the angle at the centre of the earth corresponding to the part range D_1 (figure 4) we have

$$D_1 = R\chi = R\int \frac{R_0 \sin i_0 \, dr}{r\sqrt{(\mu^2 r^2 - R_0^2 \sin^2 i_0)}}, \qquad \ldots\ldots(6)$$

$R_0 = R + h_0$, R being the radius of the earth.

As before we set

$$\mu^2 = 2\left[\frac{z^2-a^2}{b^2-z^2}\right].$$

If R_m be the radius of curvature corresponding to the level of maximum ionisation then $z = R_m - r$. Equation (6) then becomes

$$D_1 = R\int \frac{R_0 \sin i_0 \, dz}{(R_m - z)^2 \left\{2\dfrac{z^2-a^2}{b^2-z^2} - \dfrac{R_0^2 \sin^2 i_0}{(R_m-z)^2}\right\}^{\frac{1}{2}}}. \qquad \ldots\ldots(7)$$

An explicit solution for D_1 necessitates some simplification of this integral, and in this connection it is relevant to note the magnitude of the horizontal range in which we shall be interested. In the present application of the formulae we shall only be concerned with transmission distances up to 1000 km. If we assume that the effect of earth and ionosphere curvature is of the same order of magnitude in the Sellmeyer and Lorentz cases, then it is easy to show that the curvature correction to the part range D_1 only amounts to about 15 km. for a total range of 1000 km. At the maximum range in which we shall be interested, the order of the curvature correction for the part range D_1 is thus only 1·5% of the total range. Hence, for present purposes a solution for D_1 which includes a first-order curvature correction will be adequate.

A solution of equation (7) valid to terms in $1/R_m$ is

$$D_1 = \frac{2R \,.\, R_0 \sin i_0}{R_m^2 (2+s_0^2)^{\frac{1}{2}}} \left\{ (1+2\epsilon^2)^{\frac{1}{2}} \Big[F(k,\chi) - E(k,\chi) \Big]_\alpha^\beta + \frac{3\sqrt{2} \,.\, s_0 \epsilon^2}{[R_m c_0 (2+s_0^2)^3]^{\frac{1}{2}}} \right\}, \qquad \ldots\ldots(8)$$

where

$$\alpha = \sin^{-1}\sqrt{\frac{2+s_0^2}{3}}, \qquad k^2 = \frac{6\epsilon^2}{(2+s_0^2)(1+2\epsilon^2)},$$

$$\beta \simeq \frac{\pi}{2}, \qquad c_0^2 = 1 - \frac{2\epsilon^2(1-s_0^2)}{2+s_0^2},$$

$$\epsilon = f/f^0, \qquad s_0 = \frac{R_0 \sin i_0}{R_m}.$$

The derivation of this solution for D_1 is given in a note at the end of the paper.

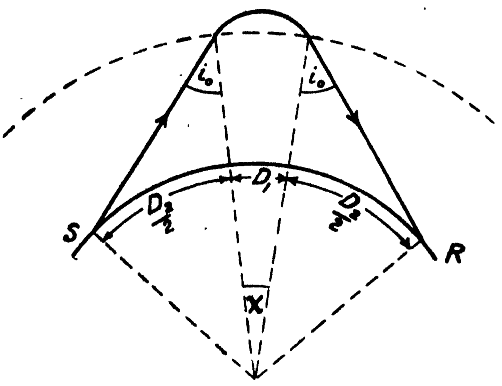


Figure 4.

The portion of the range corresponding to that part of the ray trajectory which is out of the ionosphere can be obtained from

$$D_2 = 2R\chi_2,$$

where χ_2 is given in

$$\chi_2 = \sin^{-1}\frac{[h_0 + R(1-\cos\chi_2)]\tan i_0}{R}. \qquad \ldots\ldots(9)$$

Equations (8) and (9) give the total range D, and maximum usable frequencies can now be calculated in the manner noted previously.

(e) *Case of curved earth and ionosphere. Sellmeyer formula*

In this case the solution for any type of parabolic layer has been given by Appleton and Beynon (1940, 1947), and details of the method have been described elsewhere.

Figure 5 shows M.U.F. factor curves for typical conditions for regions E and F_2 calculated from formulae (8) and (9) above. Corresponding curves for the Sellmeyer formula are also included in this figure. Table 1 gives the ratios of the M.U.F. factor calculated from the Lorentz formula to that calculated from the Sellmeyer formula for a series of distances and for reflections from region F_2. Column (2) in table 1 shows ratios calculated from the Lorentz and Sellmeyer formulae for a parabolic layer, with the constants $y_m = 100$ km. and $h_0 = 200$ km.

The values shown in column (3) compare M.U.F. factors for a layer with $y_m = 70$ km., $h_0 = 230$ km. in the Lorentz case, with M.U.F. factors for a layer with $y_m = 65$ km., $h_0 = 227$ km. in the Sellmeyer case (*vide* § 3 (c)). For purposes of comparison, ratios deduced from the work of Smith (1941) are also included in the table.

Table 1

Distance (km.)	Region F_2		
	M.U.F. factor (Lorentz formula) / M.U.F. factor (Sellmeyer formula)		Ratio of M.U.F. factors (Smith)
	$y_m = 100$ km., $h_0 = 200$ km. (S & L)	$y_m = 70$ km., $h_0 = 230$ km. (L) $y_m = 65$ km., $h_0 = 227$ km. (S)	
500	1·01	1·01	1·01
700	1·06	1·06	1·04
1000	1·13	1·13	1·12
1500	1·19	1·17	1·15
2000	1·20	1·18	1·16

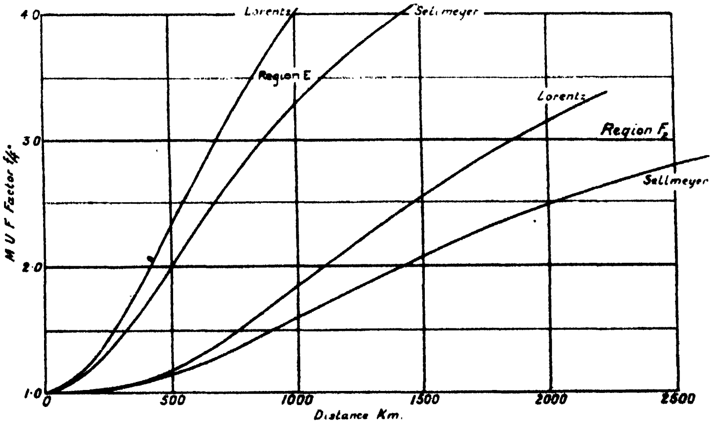


Figure 5. M.U.F. factor curves. Curved earth and ionosphere.

For any given reflecting region the magnitude of these ratios depends to a small extent on the particular values which are assumed for the layer parameters y_m and h_0, but it will be noted that the present calculations are in substantial agreement with those of Smith. At 700 km. we should expect the M.U.F. factor calculated from the Lorentz formula to exceed that calculated from the Sellmeyer formula by about 15% for region E and by 4% to 6% for region F_2. At 1000 km. the difference amounts to 17% for region E and to 13% for region F_2. The divergence between the results of Ratcliffe and of Smith appears to have arisen because Ratcliffe's calculations were confined to the shorter distances at which the differences between values calculated from the two dispersion formulae for region F_2 are certainly very small.

§4. COMPARISON WITH EXPERIMENTAL RESULTS

Reference to the curves of figure 5 shows that the divergence between the two theories is most marked in the M.U.F. factor curves for region E. Accurate measurements of the M.U.F. of region E should, therefore, provide a more sensitive test for deciding between the two theories than similar measurements on region F_2. In practice, however, measurements of the M.U.F. of the normal region E are seldom possible because, in this latitude at least, oblique-incidence reflections from normal region E are often masked by abnormal region-E reflections. The extremely variable nature of abnormal region-E reflections both at normal and at oblique incidence makes any accurate measure of the M.U.F. factor for this region extremely difficult, if not impossible. The experimental results are therefore only concerned with measurement of the maximum usable frequency of region F_2.

In two recent series of experiments, details of which are in course of publication, measurements of the maximum usable frequency for region F_2 have been made over distances of 1000 km. and 700 km. respectively, and we now consider the results of these experiments in relation to the foregoing analysis.

(a) *Observations on Zeesen broadcasting senders*

In these experiments observations were made on the signal received at Slough from the Zeesen high-frequency broadcasting station situated approximately 1000 km due east of the receiving site. Vertical-incidence ionospheric data were available only at the receiving end of the transmission path, and conditions at the mid-point were deduced from these observations. Values of the maximum usable frequency were estimated from the times of appearance and disappearance of the ray signal at sunrise and sunset. From a group of thirty-two observations it was concluded that the mean calculated value of the M.U.F. was less than that observed by 3·2%. It was realised, however, that the accuracy of these measurements was limited because of the lack of accurate ionospheric data at the mid-point of the trajectory, and much of the discrepancy seemed due to this cause. From the results of a group of twenty-one observations which were considered most reliable, it was deduced that the mean difference between observed and calculated M.U.F.s did not exceed about 2%. Now from table 1 it will be noted that for a sender-receiver distance of 1000 km. we should expect a difference of 12 to 13% between maximum usable frequencies calculated from the Lorentz and Sellmeyer formulae respectively. The fact that, in actual practice, agreement at least to within about 2% was obtained with values calculated on the Sellmeyer formula, indicates that this is the formula which is applicable to the ionosphere.*

(b) *Actual measurement of the maximum usable frequency using pulse technique*

In these experiments a large number of accurate observations of the maximum usable frequency over a sender-receiver distance of 715 km. were made using pulse technique. Vertical-incidence data obtained simultaneously with the oblique observations were available at both ends of the transmission path. The frequency radiated by the pulse sender could be varied at will and maximum usable frequency measurements made at any time. The use of pulse technique

---

* The results of a subsequent series of similar measurents over a distance of 2500 km. are also consistent with this conclusion.

enabled the maximum usable frequency of either the ordinary or extraordinary component to be actually observed visually or photographically recorded, and this, together with the fact that adequate ionospheric data for both ends were available, gives these results considerably greater accuracy than those quoted above.

Here again the observed maximum usable frequency was compared with that calculated from the vertical incidence ionospheric data on the basis of the Sellmeyer formula. The calculated value was taken to be the arithmetic mean of that calculated from the prevailing ionospheric conditions at the two ends of the path.

For a series of 110 measurements it was found that 50 observations showed an average error of $+3{\cdot}8\%$, 56 observations an average error of $-3{\cdot}75\%$ and 4 observations showed negligible error. Occasionally the divergence was much in excess of these average figures, but positive and negative discrepancies were equally frequent, and it is fairly certain that these errors were due to inaccurate knowledge of conditions at the mid-point of the trajectory. From the results of these 110 experiments it was found that the mean of all the calculated values agreed with the mean observed value to within about $-0{\cdot}2\%$. For this distance of transmission the Sellmeyer and Lorentz formulae yield calculated M.U.F.S differing by about 6% (table 1). The divergence from the Sellmeyer calculated value is thus approximately only 1/30th of what it should be if the Lorentz formula is applicable. The extremely small divergence between experimental values and the values calculated on the basis of a Sellmeyer formula again agrees with the conclusion that this is the type of formula which is valid for the ionosphere.

It may be noted that in both series of experiments the calculated value is slightly less than that observed, but it is doubtful whether this difference is significantly different from zero. Further experimental measurements will be necessary before this point can be settled. In any case the divergence is less than one-fifth of what is required if the Lorentz formula is applicable.

§ 5. CONCLUSIONS

The results of two series of measurements on the maximum usable frequency over distances of 1000 km. and 700 km. are in very close agreement with values calculated in the basis of the Sellmeyer dispersion formula. The results are in agreement with those given earlier by Smith and with the theoretical conclusions of Darwin.

APPENDIX

Note on calculation of part-range D_1 (Lorentz formula. Curved ionosphere)

We have

$$D_1 = R\chi = \frac{2R\,R_0 \sin i_0}{R_m^2} \int_{z=z_1}^{z=1} dz \left(1 - \frac{z}{R_m}\right)^{-2} \left\{2 \,.\, \frac{z^2 - a^2}{b^2 - z^2} - \frac{R_0^2 \sin^2 i_0}{R_m^2}\left(1 - \frac{z}{R_m}\right)^{-2}\right\}^{-\frac{1}{2}}. \qquad \ldots\ldots(10)$$

To terms in $1/R_m$ this can be written

$$D_1 = \frac{2R\,R_0 \sin i_0}{(2+s_0^2)^{\frac{1}{2}} R_m^2} \left\{ \int_{z_1}^{1} \left(\frac{b^2 - z^2}{z^2 - c_0^2}\right)^{\frac{1}{2}} dz + \frac{1}{R_m}\left[\int_{z_1}^{1} \left(\frac{b^2 - z^2}{z^2 - c_0^2}\right)^{\frac{1}{2}} . \, dz^2 \right.\right.$$
$$\left.\left. + \frac{s_0^2}{2(2+s_0^2)} \int_{z_1}^{1} \left(\frac{b^2 - z^2}{z^2 - c_0^2}\right)^{3/2} . \, dz^2 \right]\right\}, \qquad \ldots\ldots(11)$$

where

$$s_0=\frac{R_0\sin i_0}{R_m},\qquad c_0^2=\frac{b^2s_0^2+2a^2}{2+s_0^2},$$

and z_1 is given by

$$\frac{z_1^2-c_0^2}{b^2-z_1^2}=\frac{2s_0^2z_1}{R_m(2+s_0^2)}.$$

The three integrals in equation (11) can readily be evaluated. An expression for D_1 accurate to terms in $1/R_m$ is then found to be

$$D_1=\frac{2R\,R_0\sin i_0}{R_m^2[2+s_0^2]^{\frac{1}{2}}}\left\{(1+2\epsilon^2)^{\frac{1}{2}}\Big[F(k,\chi)-E(k,\chi)\Big]_\alpha^\beta\right.$$
$$\left.+\frac{6\epsilon^2}{R_m(2+s_0^2)}\left[\frac{\theta}{2}\left(\frac{4-s_0^2}{2+s_0^2}\right)+\frac{\sin 2\theta}{4}\left(\frac{4+s_0^2}{2+s_0^2}\right)-\frac{s_0^2}{2+s_0^2}\cot\theta\right]_{\theta_1}^{\theta_2}\right\},\qquad\ldots\ldots(12)$$

where

$$\alpha=\sin^{-1}\sqrt{\frac{2+s_0^2}{3}},$$

$$\beta=\sin^{-1}\sqrt{\frac{(1+2\epsilon^2-z_1^2)(2+s_0^2)}{6\epsilon^2}}\simeq\frac{\pi}{2},$$

$$\theta_1=\sin^{-1}\sqrt{\frac{z_1^2-c_0^2}{b^2-c_0^2}}\simeq\sin^{-1}\sqrt{\frac{2s_0^2\,c_0}{R_m(2+s_0^2)}},$$

$$\theta_2=\sin^{-1}\sqrt{\frac{1-s_0^2}{3}}.$$

$$k^2=\frac{6\epsilon^2}{(2+s_0^2)(1+2\epsilon^2)}.$$

It will be noted that θ_1 is very small and the principal term in $1/R_m$ in equation (12) is that containing $\cot\theta_1$.

Now

$$\cot\theta_1=\sqrt{\frac{b^2-z_1^2}{z_1^2-c_0^2}}=\sqrt{\frac{R_m(2+s_0^2)}{2s_0^2\,z_1}}\simeq\sqrt{\frac{R_m(2+s_0^2)}{2s_0^2\,c_0}}.$$

Substitution in (12) thus gives

$$D_1=\frac{2R\,R_0\sin i_0}{R_m^2(2+s_0^2)^{\frac{1}{2}}}\left\{(1+2\epsilon^2)^{\frac{1}{2}}\Big[F(k,\chi)-E(k,\chi)\Big]_\alpha^\beta+3s_0\epsilon^2\sqrt{\frac{2}{R_mc_0(2+s_0^2)^3}}\right\}.\qquad\ldots\ldots(13)$$

An examination of the magnitude of the term of order $1/R_m$ in either equation (12) or (13) shows that the error in D_1 introduced by neglecting this term is very small. It is found that for a layer of semi-thickness 100 km., terms in $1/R_m$ in equation (12) contribute 2 km. to a total range of 700 km., 5 km. to a total range of 1000 km. and 18 km. to a total range of 2000 km. For a thinner layer the contribution of these terms is proportionately less. It is thus to be expected that values of the total range, involving the part-range D_1 calculated from equation (13), should be accurate to better than 0·1% at all distances. Furthermore, the error involved in neglecting terms of order $1/R_m$ and using the simple formula

$$D_1 = \frac{2R\,R_0 \sin i_0}{R_m^2(2+s_0^2)^{\frac{1}{2}}}(1+2\epsilon^2)^{\frac{1}{2}}\Big[F(k,\chi)-E(k,\chi)\Big]_\alpha^\beta \qquad \ldots\ldots(14)$$

is approximately only 0·5% for a total distance of about 1000 km.

ACKNOWLEDGMENT

The work described in this paper was carried out as part of the programme of the Radio Research Board and the results are published by permission of the Department of Scientific and Industrial Research.

REFERENCES

Appleton, E. V. and Beynon, W. J. G., 1940. "The Application of Ionospheric Data to Radio-Communication Problems."—Part I. *Proc. Phys. Soc.*, **52**, 518; Part II, 1947. *Ibid.*, **59**, 58.

Booker, H. G. and Berkner, L. V., 1938. "Constitution of the Ionosphere and the Lorentz Polarisation Correction." *Nature, Lond.*, **141**, 562.

Darwin, C. G., 1934. "The Refractive Index of an Ionised Medium." *Proc. Roy. Soc.*, **146**, 17; Part II, 1943. *Ibid.*, **182**, 152.

Martyn, D. F. and Munro, G. H., 1938. "The Lorentz Polarisation Term and the Earth's Magnetic Field in the Ionosphere." *Nature, Lond.*, **141**, 159; 1939. "The Lorentz Polarisation Correction in the Ionosphere." *Terr. Mag. Atmos. Elect.*, **44**, 1.

Ratcliffe, J. A., 1939. "The Effect of the Lorentz Polarisation Term in Ionospheric Calculations." *Proc. Phys. Soc.*, **51**, 747.

Smith, N., 1941. "Oblique Incidence Transmission and the Lorentz Polarisation Term." *Bur. Stand. J. Res. Wash.*, **26**, 105.

ON THE SPECTRA OF CS AND CSe

By H. G. HOWELL,

Technical College, Bradford

MS. received 22 *February* 1945 ; *in revised form* 9 *September* 1946

ABSTRACT. **The accepted view that the main CS and CSe band systems are $^1\Pi-{}^1\Sigma$ analogous to the Fourth Positive bands is criticized, and an alternative interpretation is put forward that they are $^3\Pi-{}^1\Sigma$ transitions, possibly the counterpart of the CO Cameron bands with the $^3\Pi$ level possibly perturbed by a $^1\Pi$ level.**

§ 1. INTRODUCTION

The majority of diatomic band systems involve the ground state of the molecule, and even in cases where the nature of the lower state is uncertain (as with certain oxide emission spectra) it is probably the ground state. Consequently, when the spectrum contains a number of band systems, it is possible to acquire more extensive and accurate data for this state than for the excited states. In comparatively few instances are excited states definitely involved in more than one system, and so our knowledge of them is meagre and fragmentary compared with our well-ordered catalogue of atomic states.

It is possible to systematize the data by correlating the states of one molecule with the corresponding ones of an analogous molecule, and such a correlation should immediately reveal any missing levels for which a search can then be made. This process is simplified if rotational analyses have shown the nature of the states concerned; but if these are not available, more empirical methods must be used as a basis for the correlation. This note deals with such a method, and, as a result, the correctness of the present accepted interpretation of the spectra of CS and CSe is questioned.

§2. CORRELATION OF EXCITED STATES USING VIBRATIONAL FREQUENCIES

An attempt to correlate all states of the analogous molecules PbO, PbS, SnO and SnS was made by Howell (1936) by comparing the corresponding vibrational frequencies. He showed that although the ratio ω''/ω' (which will now be called Ω) varies irregularly from one excited state to the next for a given molecule, the same value of this ratio occurs for both oxide and sulphide molecule of the same element (Pb or Sn). This numerical agreement between states was taken to indicate a probable correspondence of state type, and it seemed to provide a reasonable method of correlation in the absence of rotational analyses. This comparison did not include the lighter molecules of this Group, viz. CO, CS, SiO, etc., as it was felt that close correspondence between the molecules of Ge, Sn and Pb (which are known to constitute a closely related group) was more likely. However, the extension to these molecules was made by Barrow and Jevons (1938) in their study of the spectrum of the group molecule SiS, and their table 7, with Ω instead of $1/\Omega$, is reproduced here as table 1, together with additional data from the later papers of Barrow (1939, 1940) and Barrow and Jevons (1940). This later work has brought in the selenides and tellurides, and it will be seen that the correspondence in Ω is wider than Howell first announced and seems sufficiently good with the exception of CS and CSe to justify the belief that the states involved are indeed corresponding ones. Barrow and Jevons consider that this correspondence does not imply that the states are all necessarily of the same theoretical type such as all $^1\Sigma$ or all $^1\Pi$ because, as they point out, whereas certain of the levels concerned are $^1\Pi$ others are accepted as being $^1\Sigma$. If this is true then there does not seem any point in making this comparison of Ω values and the correspondence found must be fortuitous and meaningless. If all the levels involved (i) have the same electron configuration, (ii) have the same type of coupling, and (iii) are free from strong perturbations, then a close correspondence between ω values can be anticipated, for condition (i) being fulfilled with these analogous molecules, the excitation of the electron to the same type of level will be accompanied by corresponding changes in the force constants from orbital to orbital, changes which will be reflected in the ω values. As for condition (ii), change of coupling is possible and it may be that the coupling of the C compounds is different from that of the Pb molecules which will probably exhibit case-c tendencies, but no change can be expected within the C and Si inner group nor in the Sn–Pb group. Consequently there is a possibility that the values of these two sub-groups refer to different levels, and the numerical agreement here found is accidental, but this point is not important for the present argument. Neglecting item (iii) for the moment, the writer believes that the agreement among the Ω values is significant and does indicate similarity of term type (possibly modified by a coupling change). Consequently any discrepancy such as occurs with both CS and CSe indicates that corresponding levels are not concerned, and also that accepted views on term type which conflict with this opinion should be carefully scrutinized. The cases of CS and CSe form the subject of this note, and the molecules PbO etc., having apparently all $^1\Sigma$ states, will be dealt with in another paper.

§ 3. ELECTRON LEVELS OF CS AND CSe

In table 1 the only serious deviations from the average Ω value of 1·45 are those of CS and CSe, which are 1·20 and 1·25 respectively. The CS system has been accepted as $^1\Pi-^1\Sigma$ (analogous to the Fourth Positive system of CO) chiefly as a result of the rotational analysis by Crawford and Shurcliff (1934). Examination of their paper immediately reveals certain weaknesses in their interpretation and provides additional stimulus to finding an alternative :

Table 1. Ω values of Group IV oxides, etc.

	O	S	Se	Te
C	1·43	1·20	1·25	—
Si	1·46	1·46	1·44	1·43
Ge	1·51	1·53	1·49	1·47
Sn	1·41	1·47	1·52	1·49
Pb	1·45	1·52	1·50	1·46

(1) The writers themselves are uneasy about the number of disturbing anomalies in their presumed $^1\Pi-^1\Sigma$ systems, e.g. under high dispersion "practically every band is found to possess double *R*-branch heads, displaced origins, abnormal $\nu_{\text{head}}-\nu_{\text{origin}}$ values and irregular branches." Two sets of *P*, *Q* and *R* branches are also noted.

(2) A scrutiny of their published spectrogram of the 0,1 band reveals other branches apparently overlooked by them.

(3) In addition to the main system, another briefer one with $\nu_e \sim 40000\,\text{cm}^{-1}$. is analysed as $^1\Sigma-^1\Sigma$ with the lower-state ω identical with ω_e'', thus making it almost certain that the ground state is involved. Yet the B'' value of this state differs by 25% from that obtained from the main system.

They interpret the main system as a $^1\Pi-^1\Sigma$, with the upper state perturbed by a $^3\Pi$ state. If this is so, then under condition (iii) it would be possible to account for the discordant Ω values. However, the large extent of the anomalous features spread over practically all the bands indicates such a violent perturbation that the writer has sought for any possible alternative explanation. A $^3\Pi-^1\Sigma$ transition is immediately suggested, in which case the system would be analogous to the Cameron bands of CO. This would possibly account for all the observed complexities and certainly for the increased number of branches. Indeed, comparison of the structure of the Cameron bands as analysed by Gerö, Herzberg and Schmid (1937) with that of CS reveals quite a close resemblance. Also the upper $^3\Pi$ state of CO has an Ω value of 1·25 against the 1·20 of CS. The main objection to this view is one of intensity—such a transition should be much weaker than the corresponding $^1\Pi - {}^1\Sigma$, but in this connection the experiments of Knauss and Cotton (1931) on the intensities of the CO bands are very illuminating. They observed that whereas high-pressure excitation conditions favoured the production of the Fourth Positive, low-pressure conditions tended to suppress them whilst increasing the intensity of the Cameron system.

This selective effect of pressure on the singlet and triplet systems can be expected to occur also in CS and CSe, with the triplet system becoming relatively stronger in the direction CO—CS—CSe owing to the increase in molecular weight. Consequently, such a $^3\Pi - {}^1\Sigma$ transition should have a higher probability with CS and CSe than with CO if the conditions are suitable. Now CS bands can be obtained under high- or low-pressure conditions, and Martin (1913) observed that when they were produced in a carbon arc the most intense bands were around 2579 A., and these were suppressed in the Geissler discharge which is used for the normal CS system. Crawford and Shurcliff actually state that very low pressures were necessary in their work in order to eliminate the extensive S_2 spectrum, and also that higher pressures decreased their intensity. It seems, therefore, that the conditions eminently suit the production of a $^3\Pi - {}^1\Sigma$ system.

The intensity objection would then appear to have lost much of its strength. It can also be pointed out that the extent of the so-called perturbation of the $^1\Pi$ means a strong interaction of the $^3\Pi$and $^1\Pi$ which is violating the singlet-triplet prohibition just as much as in a $^3\Pi - {}^1\Sigma$ transition and that any objection to the intensity of the $^3\Pi - {}^1\Sigma$ system applies equally well to this perturbation.

The presence of the 2579-A. bands (which appear to have been forgotten by subsequent workers) would indicate the close presence of another state and possible perturbation. Consequently the correct interpretation of the so-called $^1\Pi - {}^1\Sigma$ system should be among the following:—(*a*) a $^3\Pi - {}^1\Sigma$ without perturbation; (*b*) $^3\Pi - {}^1\Sigma$ with the upper state perturbed by $^1\Pi$; (*c*) $^1\Pi - {}^1\Sigma$ with the upper state perturbed by $^3\Pi$.

If (*b*), the upper state of the 2579-A. bands can be identified as the $^1\Pi$. From its Ω value, the only known excited state of CSe is analogous to that of CS just discussed, but no rotational analysis is available. Examination of Barrow's paper (1940) on the vibrational analysis shows that in attempting to interpret the system as $^1\Pi - {}^1\Sigma$, the same difficulties are encountered as with CS. For example, it is necessary to introduce a cross term in $u'u''$ to account for the systematic deviations of ΔG values, due apparently to the variation in interval between *R* head and origin which should surely be small for this molecule as well as for CS. Even with this additional term it is also found necessary to postulate a vibrational perturbation for $v'=1$, and finally a few bands cannot be definitely allocated. It would seem that a transition of the type $^3\Pi - {}^1\Sigma$ would account for certain of these difficulties.

§4. CONCLUSION

It would appear that there is room for more experimental investigation of these molecules, particularly of the 2579-A. bands of CS. Also, even if the alteration in the upper state type of these molecules, suggested here, is incorrect, the difference in the B'' values found by Crawford and Shurcliff still remains to be accounted for.

REFERENCES

Barrow, R. F., 1939. *Proc. Phys. Soc.*, **51**, 269.
Barrow, R. F., 1940. *Proc. Phys. Soc.*, **52**, 380.
Barrow, R. F. and Jevons, W., 1938. *Proc. Roy. Soc.*, A, **169**, 45.
Barrow, R. F. and Jevons, W., 1940. *Proc. Phys. Soc.*, **52**, 534.
Crawford, F. H. and Shurcliff, W. A., 1934. *Phys. Rev.*, **45**, 860.
Gerö, L., Herzberg, G. and Schmid, R., 1937. *Phys. Rev.*, **52**, 467.
Howell, H. G., 1936. *Proc. Roy. Soc.*, A, **153**, 683.
Knauss, H. P. and Cotton, J. C., 1931. *Phys. Rev.*, **38.**, 1190.
Martin, L. C., 1913. *Proc. Roy. Soc.*, A, **89**, 127.

A SIMPLE OPTICAL MODEL DEMONSTRATING THE PRINCIPLE OF THE BRAGG X-RAY SPECTROMETER

By F. A. B. WARD,
The Science Museum, London S.W.7

MS. received 22 *February* 1946; *demonstrated* 5 *July* 1946

ABSTRACT. The model illustrates by an optical analogy the fact that a monochromatic beam of x rays will be reflected by a single crystal only when the crystal is set at a particular angle to the incident beam such that reflections from successive sheets rich in atoms are in phase and so give an interference maximum. The direction of the reflected beam can also be determined.

---

§ 1. GENERAL DESCRIPTION OF THE MODEL

The model is designed for exhibition in the Science Museum to illustrate the principle of the Bragg x-ray spectrometer. It can be constructed, however, of such simple materials and at such a low cost that it should be of general use for teaching the fundamental principle of x-ray analysis of crystals.

As shown in the photograph, the incident beam of x rays is represented by a sheet of corrugated iron, 18 × 12 in. The spacing between successive crests is 3 in., so that six complete waves can be seen. It is advantageous, though not essential, to paint the system black and to paint a white strip ¼ in. wide along the crest of each wave.

The crystal to be analysed is represented by four glass plates (12 × 10 in. photographic negatives with the emulsion removed, the longer dimension being vertical). The plates are mounted one behind the other, 2 in. apart, in a simple wooden rack; an additional grooved distance piece is placed on top of the plates to ensure more accurate parallelism. The rack is provided with a plywood hood in order to cut out stray light. The rack, whose horizontal dimensions are 10¾ × 7 in., is pivoted on a central spigot, and carries two pointers indicating upon a scale of degrees.*

---

* In the Science Museum model this consists of a 12-in. celluloid protractor placed over a white ivorine disc.

In use, the corrugated iron is illuminated from above by means of an adjustable table lamp, and the observer looks into the glass plates. Images of the corrugated iron are seen by reflection from each plate, but in general the crests of the corrugations will not be in phase, and a confused picture is seen. If, however, the plate-rack is rotated until the relation $n\lambda = 2d \sin \theta$ is satisfied, where λ is the distance between corrugations and d is the spacing of the plates, then all the reflected images of the crests will be in phase and the images will coalesce, giving a simple image as of a single corrugated sheet. θ is then the angle at which a reflected x-ray beam would be obtained.

The principle is thus illustrated, but for more accurate measurements it is advisable to replace the corrugated iron by a flat wooden board of similar dimensions, black but with white lines ¼ in. wide, spaced at 3-in. intervals. These do not so obviously represent waves, but they can be more accurately spaced than the crests of the corrugated iron, which depart slightly from a true sine-wave form. As before, the plate-rack is rotated until the successive white lines appear accurately superposed. In order to evaluate θ, readings α_1 and α_2 are taken on either side of the incident beam; the angle $\alpha_2 - \alpha_1$, through which the plate-rack has been turned, is then equal to $180° - 2\theta$, so that

$$\theta = 90° - \frac{\alpha_2 - \alpha_1}{2}.$$

Having determined θ, one can either assume λ to be known and evaluate d or vice versa.

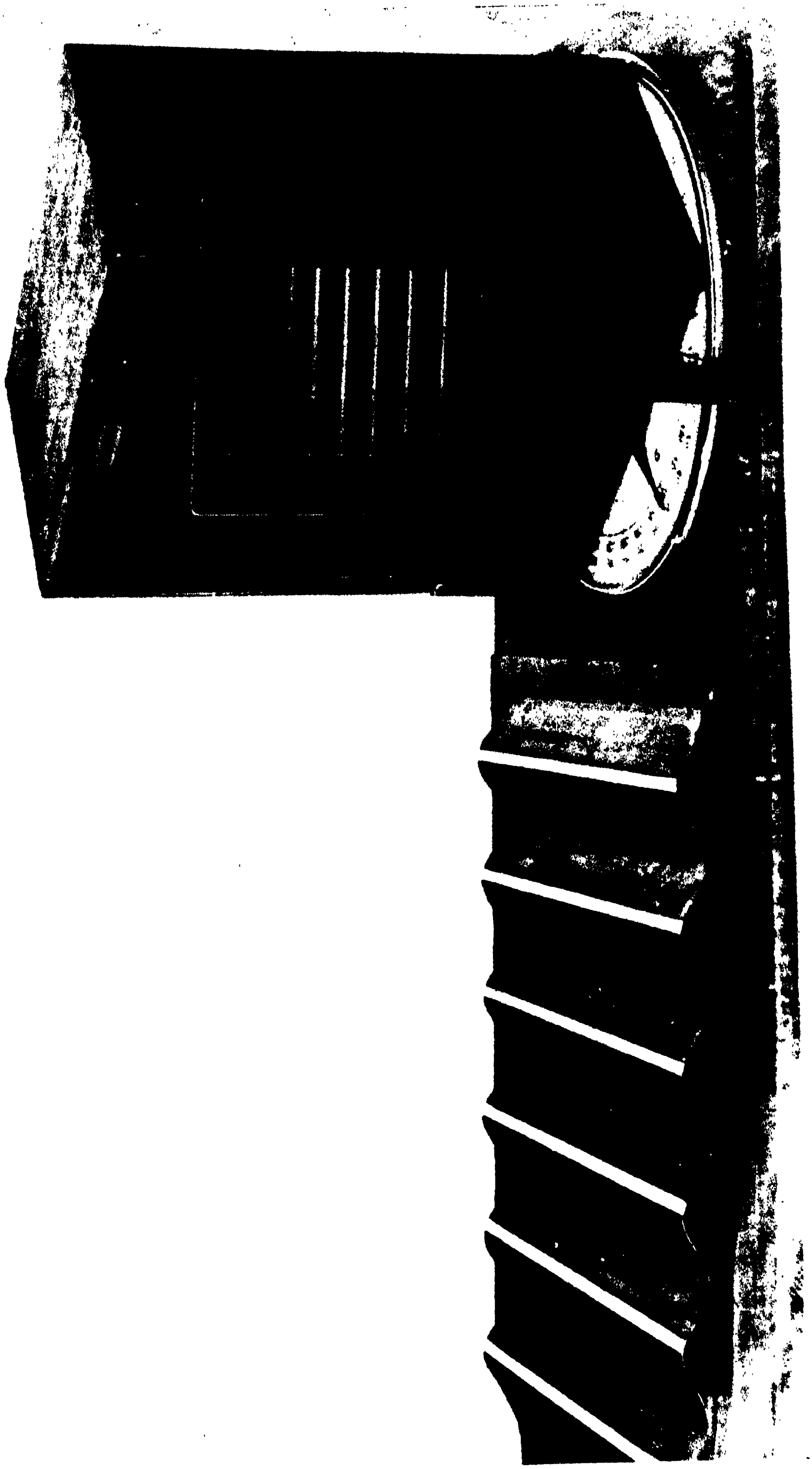
To facilitate the interchange of the corrugated iron and painted wood, the former is mounted on a wooden board, in each corner of which a ½-in. hole is bored. These holes fit over wooden dowel pins projecting from the general baseboard. Similar holes are bored in the corners of the blackened wooden board carrying the white lines.

In order to find the direction of the reflected beam, the model is provided with an extra arm, as shown in the photograph. This arm rotates about the same spigot as the plate rack; it is bent upwards and carries a horizontal straight-edge, perpendicular to the axis of rotation, and 6 in. above the baseboard. In use, this arm is rotated until the straight-edge is seen to be parallel to the images of the wave-crests; the edge is then perpendicular to the reflected beam, whose direction can be read off by means of a pointer mounted on the arm and indicating on the circular scale. It will be seen that a simple wire view-finder is mounted above the straight-edge.

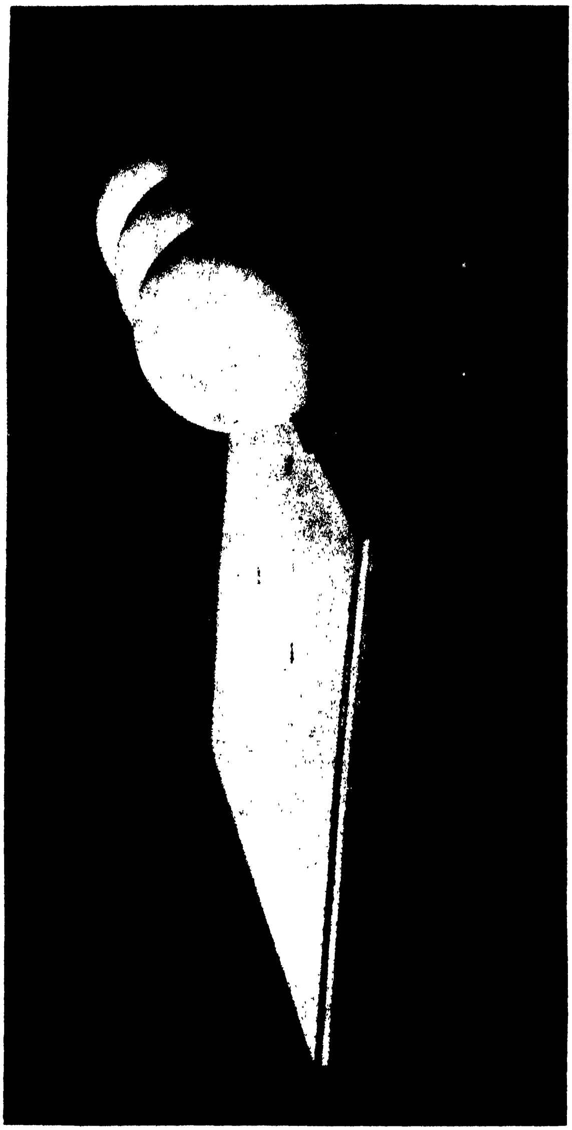
§ 2. LIMITATIONS OF THE MODEL

The finite thickness of the glass plates causes a doubling of each image, but, provided that the thickness of each plate is a small fraction of the spacing of the plates, this does not interfere seriously with the efficiency of the model; it was for this reason, however, that it was made on such a relatively large scale.

It would appear possible to design a smaller model, using corrugated cardboard instead of corrugated iron, and glass plates of quarter-plate size, one side of each plate being "coated" to reduce reflection in the manner now employed in various optical instruments and camera lenses. The coating would, of course,



A simple optical model demonstrating the principle of the Bragg x-ray spectrometer.



A single unit of the uranium chain reaction model.

produce its maximum effect only at a particular angle of incidence, and it would be preferable to illuminate the incident "wave" with monochromatic or approximately monochromatic light.

ACKNOWLEDGMENT

I wish to thank Dr. H. Shaw, Director of the Science Museum, for permission to publish this paper.

A MECHANICAL MODEL ILLUSTRATING THE URANIUM CHAIN REACTION

By F. A. B. WARD,
The Science Museum, London S.W.7

Demonstrated 5 *July* 1946; *MS. received* 16 *October* 1946

ABSTRACT. The model consists of 30 units, each representing the nucleus of a uranium atom. Each unit consists of two portions, held together by a form of catch which, when released, allows the two portions to fly apart, while at the same time three table-tennis balls, representing neutrons, are projected upwards to a height of two to three feet.

To demonstrate the model, the 30 units are placed in rows in a glass showcase measuring 6 × 3 × 3 ft. high. A single table-tennis ball is dropped upon the trigger of one of the units; this unit immediately disintegrates, projecting its three " neutrons ", which impinge upon neighbouring units, causing them to break up and initiating a " chain reaction " which spreads until a majority of the units have disintegrated, the whole process occupying a few seconds.

The units are designed and their layout planned so as to secure maximum efficiency.

§1. GENERAL DESCRIPTION

The model was designed for demonstration in the Atomic Energy Exhibition which opened at the Science Museum on 14 February 1946. It is intended to illustrate the typical uranium chain reaction which has been utilized for the release of atomic energy. The model consists of a number of units, each representing the nucleus of an atom of ^{235}U or of plutonium. In designing the units, experiments were first tried with a breakback mouse-trap. These showed promising results, and from them the type of unit now employed was evolved. The detailed design of this unit, an example of which is shown in the photograph, is due to Mr. G. H. C. Jones, Deputy Foreman of the Metal-Working Shop at the Science Museum.

Each unit consists of two portions, shown in figures 1 and 2 respectively, which can be coupled together in a manner to be described. The first of these portions, shown in plan and elevation in figure 1, consists of two blocks of wood, A and B, hinged together at C, and fitted with short projecting steel pins D and E. The blocks are normally kept apart by the helical springs F and G, their separation being limited by the head of the retaining screw Z. To couple the two portions of a unit together, ready for operation, the upper block A is pressed down by hand, and the two pins D and E are inserted into the holes H and I in the other portion of the

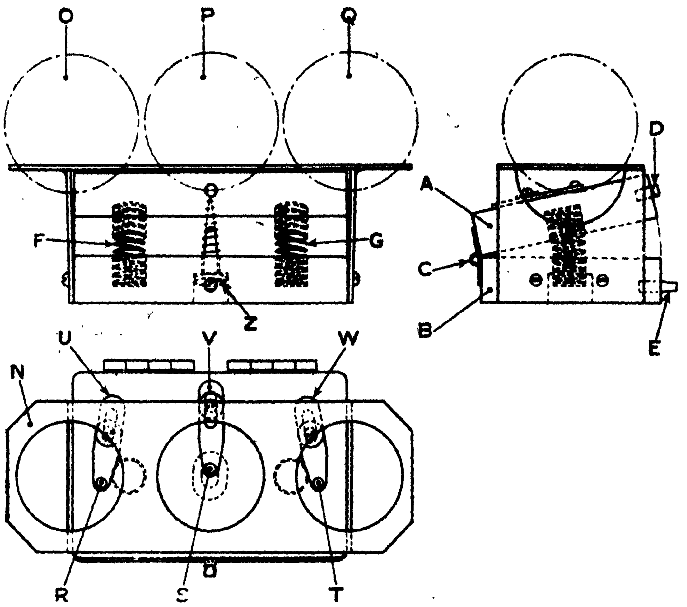


Figure 1.

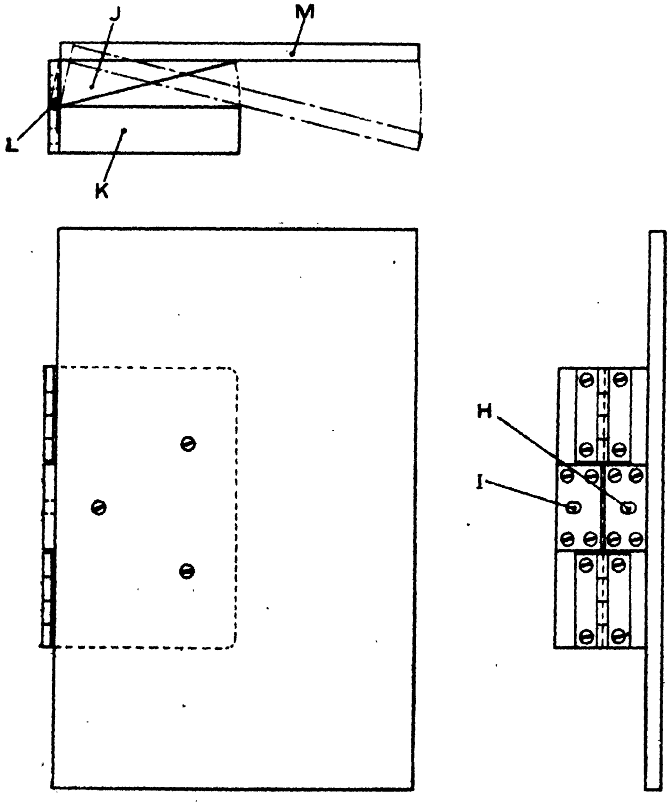


Figure 2.

unit. This second portion consists of two wooden blocks, J and K, hinged at L. A rectangular "target" M, of thin plywood, is screwed to the upper block J and, as shown in the photograph, lies approximately in a horizontal plane when the unit is "set".

When a table-tennis ball falls upon this target, it pushes it down from the position shown by the full lines in figure 2 into that shown by the chain-dotted lines, and releases the upper pin D from the hole H. The helical springs F and G then force the upper block A of the first portion of the unit smartly upwards, until it is stopped by the head of the retaining screw Z. As it rises, it strikes the three table-tennis balls O, P and Q, which rest upon circular holes in a brass frame N which is screwed to block B by means of side flanges. In the unit as first designed, the balls were struck directly by the upper surface of A but, in order to give some control over the flight of the balls, three adjustable strikers were later fitted. These consist of small steel rivet-heads R, S and T mounted on adjustable slotted brass strips U, V and W. In the absence of the strikers, the balls are projected at an angle of about 20–30° from the vertical, with speed sufficient to carry them up to a height of four to five feet, so that they fall several feet away from the unit. The strikers can, however, be adjusted so as to project the balls more nearly vertically by means of a more glancing blow which also gives a lower speed of projection. In practice, the strikers are arranged so that the balls are projected up to a height of two to three feet, and fall about six to twelve inches away from the unit.

In setting up a unit, it is found best to push the projecting pins of the trigger fully home into their sockets and then to ease them a little, so as to make the trigger more sensitive.

To demonstrate the chain reaction, the units are lined up in row on the baseboard of a standard museum showcase whose interior dimensions are 6 x 3 x 3 ft. high. The lay-out adopted is shown in figure 3.

The first row contains a single unit, the second row four units, and the remaining rows five each. In the first four rows, all units are arranged so as to throw the balls forward to the next row; but in later rows some units are arranged to throw backwards, in order to set off any earlier units which may have escaped being struck. *All* the units in the last row are arranged to throw backwards, for obvious reasons.

The reaction is initiated by dropping a single table-tennis ball upon the unit in the first row. The releasing mechanism is extremely simple, and is roughly illustrated in figure 4, which shows pictorially the general course of the reaction. The ball to be dropped is placed inside a short cylindrical brass tube, and is normally prevented from falling by a transverse diametral pin passing through holes in the sides of the cylinder. The pin can be withdrawn by pulling a string attached to one of its ends. The ball then falls upon the target of the unit in the first row; this unit explodes, projecting its three "neutrons" upwards and to the right (see figure 4); some of these "neutrons" impinge upon the targets of other units, which also blow up, and the "chain reaction" is established. If a relatively large number of hits happen to be scored in the first few rows, the action becomes almost "explosive" in type, and lasts only about three seconds, but, if the early number of hits is smaller, the release of energy proceeds more steadily, lasting 5 to 10 seconds.

§2. NUMERICAL REQUIREMENTS FOR THE MAINTENANCE OF THE "CHAIN REACTION"

It is of interest to determine the conditions limiting the development of the chain reaction.

The target area of each unit measures 6 × 4 in. and, if the unit is properly adjusted, any ball striking this area will produce a disintegration. If the ball strikes any other part of a unit—such as the table-tennis balls or their supporting frame—it makes an inelastic collision and is lost to the reaction. Each unit has accordingly what we may term a "dead area", which measures a little less than 6 × 4 in.

Figure 3.

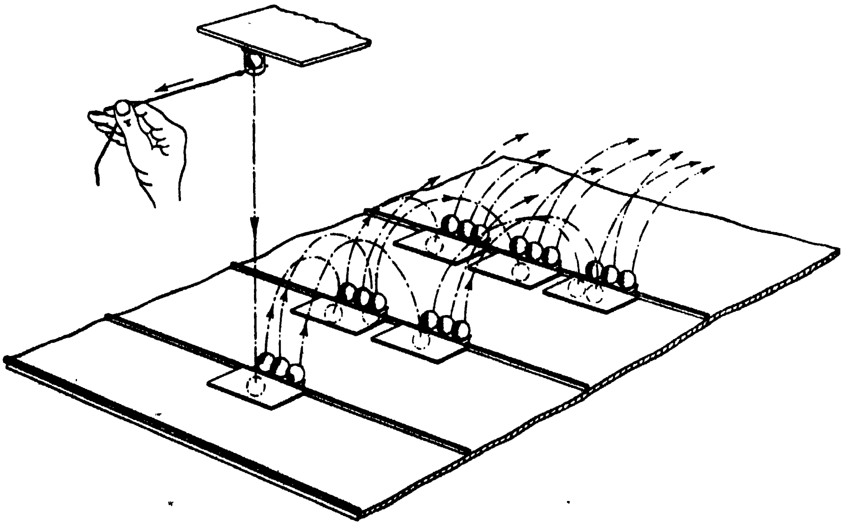


Figure 4.

In a disintegration, the two halves of a unit fly apart two to three inches before coming to rest, and hence the units cannot be placed too close together, or they will set each other off by direct collision, and the analogy with the uranium reaction will be spoiled. With the layout shown in figure 3, the total floor space occupied is 70–80 sq. in. per unit. Of the space occupied by the model, therefore, roughly one-third represents live target area, one-third " dead " area of the units themselves, and one-third area not covered by units. It can easily be seen that, if this last area is non-reflecting to table-tennis balls, a chain reaction will be difficult to maintain. For consider the fate of the three balls projected from the first unit disintegrated. The chance that *one* of these will score a miss is 2/3. The chance that all three will miss is $(2/3)^3$ or 8/27. In 8 cases out of 27, therefore, we can expect that the reaction will not even start.

If, however, the floor space not occupied by units is totally reflecting, so that the balls rebound from it to their original height, the chance that one of the balls projected from the first unit will score a miss is 1/2, for it is equally probable that when the ball strikes a unit it will strike the " live " or the " dead " part of it. In this case, therefore, the chance of all three balls missing, and the reaction failing to start, is reduced to 1 in 8, with a corresponding increase in the probability that, once started, it will develop.

In the actual model, the baseboard of the case containing the model is of painted wood, from which the balls rebound to a height of about two-thirds of that from which they fall. This is sufficiently close to total reflection to enable the reaction to proceed satisfactorily. If the ball is effective after one bounce, but not after two, it can be shown that the probability of all three balls from the first unit missing, and the reaction failing to start, is $(5/9)^3$ or approximately 1/6. In practice, the results are more favourable than either the figure 1/6 or 1/8 would suggest, and a complete failure is very rare indeed. This is mainly because the strikers of the first unit can be so adjusted that the balls have a very good chance of falling on the targets of the second row—they can, in fact, be definitely " aimed " at these targets; it is also probable that the " dead " area of each unit has been over-estimated in the above analysis, and furthermore is not entirely " dead ".

In a single demonstration, the average number of units disintegrated is about 26 out of the total of 30. As the reaction proceeds, the floor of the case becomes littered with disintegrated units, so that the proportion of " dead " area begins to increase rapidly, targets of disintegrated units being now " dead " area. One cannot therefore normally expect to disintegrate all 30 units, though it can be done on relatively rare occasions when chance is fortunate.

ACKNOWLEDGMENTS

I wish to express my thanks to Dr. H. Shaw, Director of the Science Museum, for permission to publish this paper. I am much indebted to Mr. G. H. C. Jones, who was responsible for the detailed design of the units. I also wish to thank Mr. S. H. Groom for valuable criticisms and suggestions relating to the lay-out of the units.

THE EMISSIVITY OF HOT METALS IN THE INFRA-RED

By DEREK J. PRICE,
S.W. Essex Technical College, London

Communicated by H. Lowery; *MS. received* 15 *July* 1946

ABSTRACT. A specially designed vacuum-tight, water-jacketed furnace has been used in the measurement of the emissivity of incandescent metals for wave-lengths of 1·0 to 4·5 μ. Results have been given for pure platinum, iron, molybdenum, copper and nickel. A " split cylinder " type of comparison black-body has been used and its efficiency has been compared with a spherical standard. This calibration has suggested a correction to be applied to Ives' standard of visible radiation. From the present emissivity values, and from those given by other workers, it appears that for a wave-length (usually in the near infra-red) peculiar to each metal the emissivity is constant over a large range of temperature. Such *X-points* have been observed in the case of iron and molybdenum, but not with platinum or copper.

§ 1. INTRODUCTION

The study of radiation emitted from hot objects is important for at least two reasons. On the one hand, such a study is fundamental to the practice of optical and total radiation pyrometry, and on the other, the development of modern physics has made it clear that the theory of radiation is a basic section of our knowledge of atomic and electronic mechanism. The " perfect black-body " of which the properties are described by Planck's law of radiation does not correspond with any real substance : indeed, a classical example has been cited by Coblentz (1908).

> " It is, of course, absurd to attempt to measure the temperatures of these substances (complex minerals) with an optical pyrometer. For example, the rod of oligoclase was a perfectly transparent glass and emitted no light on heating to over 1200° C., except that due to the sparking of the platinum terminals. Nevertheless, such substances as iron oxide at the same temperature would have emitted visible light, while both emit strongly in the infra-red region."

A number of attempts have been made to formulate directly the radiation laws of imperfect radiators but these have not met with much success. It is more profitable to consider the degree of imperfection of the radiator, taking as a numerical measure the ratio of the energy radiated by an imperfect and a perfect body respectively under the same conditions. This ratio, having a value between 0 and 1, defines the *emissivity* of the material. Kirchhoff's law (stating that the sum of the emissivity and reflectivity of an opaque body must be unity) acts as a connecting link between the theory of optical constants.

Hagen and Rubens (1900) at the end of the nineteenth century stimulated interest in this field by providing and testing successfully a theory of emissivity valid for long wave-lengths. Their equation $E=2\sqrt{\nu/\sigma}$ also accounted for the observed variation of emissivity with temperature since the change of σ (the electrical conductivity) with temperature was known independently for the metal used; ν is the frequency of the radiation. For wave-lengths of the order of $10\,\mu$ or more, good agreement with this theory was found, but for shorter wave-lengths it failed completely. The temperature variation, too, being derived from the Hagen–Rubens relation, broke down for shorter wave-lengths. More recent work has shown that in the visible region some metals at least appear to have a temperature coefficient opposite in direction to that predicted by the simple theory. For all metals, the magnitude of this coefficient is very much smaller for short wave-lengths than that predicted by the Hagen–Rubens relation. For those few metals for which sufficient data have been obtained, it seems that there is in the near infra-red a wave-length for which the temperature coefficient is zero over a very large range of temperatures. This remarkable occurrence of a special wave-length (here called the *X-point*) peculiar to each of these metals does not appear to have been examined critically by any previous worker, although the possible existence of such a point has been noted by Worthing (1926). On the basis of two observed values he suggested that the X-point wave-length was proportional to the melting-point of the metal. This is at variance with other known X-points. The present work has been concentrated on the provision of more accurate data for the transition region between the range of validity of the Hagen–Rubens relation and the X-point in hope of further elucidating this phenomenon.

§ 2. PREVIOUS METHODS

The main methods that have been used to measure the emissivity of metals may be divided into five categories as follows :—

(1) Evaluation from measurement of the optical constants.
(2) Determination of reflectivity.
(3) Measurement of the true and apparent temperatures of a hot substance.
(4) Measurement of the quantity of energy radiated by a hot body at a known temperature.
(5) Direct determination as the ratio of the energies radiated by the hot body and a comparison black-body under the same conditions.

The first method is theoretically superior to the others since it yields two independent quantities, n and k, of which the emissivity is only a function. Against this, however, is the difficulty of catoptric observations in the infra-red and the added complication of working at temperatures as high as was desired. The second method has been more widely used than any other on account of its simplicity. There are three objections to its use in the present case. First, the hot metal mirror will radiate energy which must be compensated for by a method similar to that used by Beekman and Oudt (1925). Secondly, at high temperatures, the development of crystal facets and the warping of the mirror both tend to interfere with specular reflection. Thirdly, as has been pointed out by Hurst, a 1% error in the measured reflectivity may, in the infra-red, be equivalent to a 20% error in the calculated emissivity. This is due of course to the high reflectivity in this region of the spectrum. The measurement of true and apparent temperature is a very convenient method for use in the visible spectrum but difficulties become almost insuperable in the infra-red. The fourth method was abandoned because of the difficulty of making temperature and absolute energy readings of a sufficiently high order of accuracy. Preliminary calculations (Price and Lowery, 1944) have shown that the results obtained for instance by Hase using this method could not be satisfactorily converted to emissivity values.

The last method is free from all of the previously noted faults and presents only one difficulty, namely the provision of a comparison black-body. This has been the subject of some ingenuity in previous work, at least four different arrangements having been used to ensure equality of temperature between the specimen and black-body. Historically, Mendenhall (1911) was the first to use a method of this nature. His specimen was in the form of an electrically heated strip of thin metal bent to the shape of a V of small angle. Due to multiple reflection, black-body conditions obtain on sighting into the opening of the V. An improvement of this was made by Hurst (1933) who used a block of metal with a V cavity cut in it. This obviated the buckling of the metal which, in Mendenhall's case, proved to be the greatest source of inaccuracy. The disadvantage of Hurst's method is that for each specimen a specially machined block must be made. Other types of comparison black-body that have been used include a hole in a tubular filament (Langmuir, 1916) and a helical filament or strip (Worthing, 1925). Neither of these devices yields a conveniently shaped source for spectroscopic observation. Another device having all the advantages of Hurst's wedge apparatus, but allowing an easy change of specimen, was used by Drecq and later investigated by Ives (1924) as a primary standard of luminous radiation. These workers used a thin sheet of platinum bent into the form of a cylinder with the edges left slightly apart so as to form an axially placed slit. The cylinder was heated ohmically by means of copper electrodes at either end, to which it was rigidly clamped. It was decided to investigate whether this device could be readily adapted to working in an inert atmosphere or a vacuum in order to observe oxidizable metals.

§ 3. EXPERIMENTAL ARRANGEMENT

The first model furnace was made in the form of a small brass box enclosing a specimen 5 cm. long and 1 cm. diameter, heated by a current of about 300 amp. obtained from a single turn secondary on a 1 kVA. mains transformer, and observed through a small window of fused silica. The performance of this trial model may be summarized as follows :—

(1) Satisfactory results may be obtained with fixed terminal blocks since the deformation produced by thermal expansion of the specimen is small.

(2) The slight deformation so produced does not noticeably affect the radiation from the black-body slit.

(3) The centre portion of the specimen is at a reasonably uniform temperature.

(4) It is advisable to work *in vacuo* or in a stationary atmosphere since the specimen temperature is extremely sensitive to draughts.

(5) The temperature of the furnace becomes quite high (300° c.) after running for a short while and this destroys the efficiency of most forms of vacuum sealing cement.

A second model furnace of more massive design was then constructed with the vacuum joints formed by surface-ground steel plates bolted together. Some preliminary results were obtained with this apparatus, but the rapid heating up of the furnace body caused so many difficulties that a third model was constructed incorporating water cooling for the furnace, the electrodes, and the transformer secondary turn. It was made in two halves: a steel base plate and a one-piece cover of fused silica incorporating a window of optical quality. The design can be seen from figure 1. The electrodes were sealed in with vacuum wax and the joint between the two halves of the furnace was kept smeared with vacuum grease. The furnace was mounted on a board attached to the transformer and the whole assembly formed a rigid unit. This apparatus was used without alteration for all subsequent work. Its completely satisfactory performance may be summarized as follows :—

(1) The water cooling was so efficient that no stray heat could be detected and ordinary paraffin wax could be used for temporary seals.

(2) A good vacuum was obtained even with specimen temperatures of 1500° c.

(3) A cylindrical specimen could be made from sheet metal and inserted in the furnace ready for emissivity readings to be taken within ten minutes.

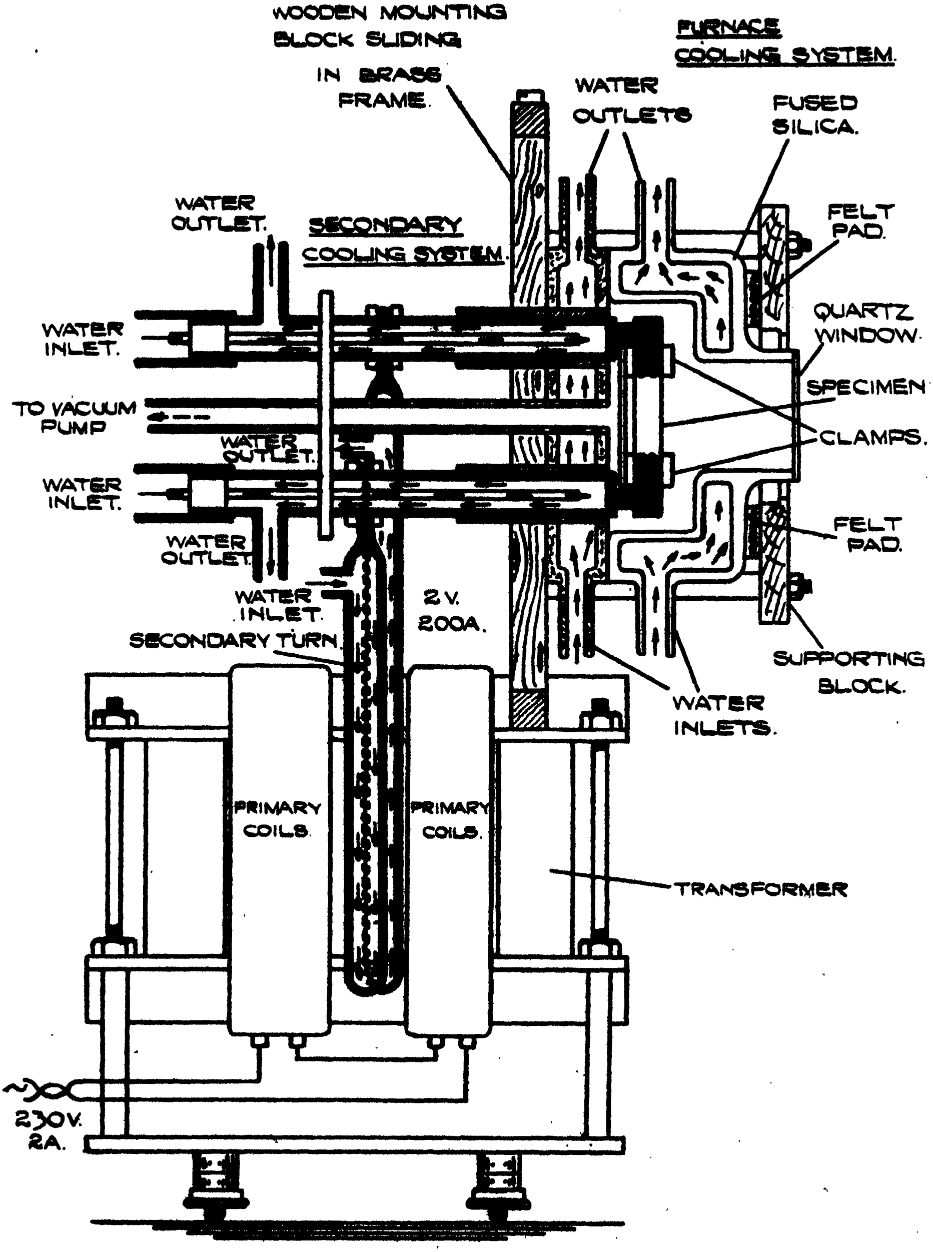


Figure 1. Emissivity furnace with auxiliary equipment.

This permitted the use of a large number of specimens in order to average out stray variations in the surfaces, etc.

A Schwartz-type thermopile mounted in air was used in conjunction with a Hilger Barfit Infra-red Spectrometer employing a rock-salt prism in a constant-deviation mounting. The thermopile was connected through a series resistance box to a Tinsley 4500 LS galvanometer. The complete arrangement was free from serious trouble both during the initial calibrations and throughout the subsequent experiments. The spectrometer and furnace, together with a large mirror for focusing the radiation from the furnace on the slit of the instrument, were mounted in a draught-tight case containing trays of calcium chloride to safeguard the rock-salt prism. A system of gearing and an observation window enabled the wave-length drum of the spectrometer to be adjusted from outside. Another window at the end of the case enabled an optical pyrometer to be sighted on the specimen inside the furnace. The focusing mirror was fitted with a slow-motion screw to rotate it so as to set either the black-body slit or the specimen surface adjacent to it on to the spectrometer slit. A series resistance and multi-tapped transformer enabled the input to the furnace transformer and therefore the temperature of the specimen to be controlled externally. A shutter controlled by a cable lead was fitted in front of the spectrometer slit. The vacuum tube from the furnace was connected by means of a two-way tap to a Hivac vacuum pump and to a cylinder of pure hydrogen (99·8 % H_2) supplied by the British Oxygen Co. By this means the furnace could be evacuated, flushed or filled with hydrogen. In later experiments an auto-transformer mains stabilizer was used to obviate fluctuations of the heating current that occurred during long series of observations.

§4. EXPERIMENTAL TECHNIQUE

The raw material for the specimens was in each case a sheet of metal about 0·2 mm. thick. After selecting a uniform area, a rectangle 5 cm. × 3 cm. was cut with a clean razor blade. This was then filed to the exact size needed for the required slit width, using a steel former. The sheet was then bent into a cylinder between two brass formers round a rod and subsequently annealed. After polishing with very fine emery (when this was necessary) the specimen was washed in alcohol, dried, and kept over calcium chloride until required. The polishing process was not often used since it was found that after a specimen had been used once at a high temperature, and removed from the furnace, the specimen surface had a very good, polished appearance due no doubt to local evaporation taking place on the scratches. It was found, however, that this "self-polishing effect" produced no measurable change in emissivity, and hence the effect of small surface scratches was disregarded. Prior to the fitting of each specimen, the electrodes and clamps were cleaned with emery cloth to ensure good contact. The specimen having been fitted, the furnace was evacuated for a few minutes to remove moisture and then the furnace was filled with hydrogen and evacuated again. After repeating this process and filling again with hydrogen, the furnace was ready for use. The furnace current was switched on, adjusted to the required value, and pyrometer readings were then taken on the specimen surface and on the black-body slit. At this stage any large buckling of the specimen could be observed and corrected by opening the furnace. This, however, was not a frequent cause of trouble. After this, emissivity readings were taken up and down the wave-length range at intervals of 0·1, 0·25 or 0·5 μ. Finally the pyrometer was read again as before. Each emissivity reading, with necessary repeats, involved nine galvanometer deflections.

The portion of the specimen to be focused was the same for all wave-lengths and was usually a strip one or two mm. distant from the black-body slit. For each emissivity reading, the zero and sensitivity of the galvanometer were adjusted so as to give maximum accuracy. The observations made with the optical pyrometer were used to calculate the average emissivity for a wave-length of $0{\cdot}65\,\mu$ by means of the pyrometer correction formula,

$$\frac{1}{T}-\frac{1}{T_{\text{app}}}=-\frac{\log E}{9880}.$$

The galvanometer readings were averaged so as to give for each wave-length a "black-body deflection" and a "specimen deflection". The ratio of these (specimen: black-body) gave the calculated emissivity at this wave-length.

§ 5. CORRECTIONS

One possible source of error, the existence of a temperature difference between the inside and the outside of the hot cylinder, has been investigated by Angell (1911) who found the effect negligible. A second source of error (that of a temperature difference round the cylinder) was studied by Ives (1924) and was also found to be negligible. A photographic examination of the present furnace confirmed this result. It must, however, be remarked that the axial temperature-gradient found with this arrangement was seen, upon micro-photometric examination, to be much larger than that reported by Ives with a specimen of similar dimensions. This error was nullified in the present work by taking the ratio of the radiation from the slit and a parallel section of specimen surface having a similar gradient, but it is probable that the use of such an arrangement as a primary standard of light would be seriously impaired by this gradient unless a much longer cylinder were used.

From the outset of the present work it was found that the emissivity values obtained were rather high not only in the infra-red but also at $0{\cdot}65\,\mu$. Since the discrepancy appeared to increase with wave-length and was approximately the same for platinum and iron, it pointed to an imperfection in the black-body. The only direct check on this by means of a standard black-body furnace proved to be unobtainable due to war-time difficulties, which made it impossible to obtain and calibrate even a suitable sub-standard. The exact source of the error was not easy to find. The effect of a finite instead of an infinitesimal slit has been discussed by Ives and found to be small provided the slit is sighted from an angle inclined a few degrees in azimuth from the normal. This condition was complied with in the present work and, moreover, the results of Ives were independently confirmed by photographic means. Another investigation by Cunnold and Milford (1934) on the black-body deficiency caused by a finite hole in the wall of a hot tube asserts that in general the error from this source is very small.

A series of experiments with slits ranging from 0·25 to 2·0 mm. in width showed very little variation, so confirming the above. A specially designed specimen having a uniformly heated central portion produced no better result, and hence it was deduced that the axial temperature gradient was not the source of error. Roughening the interior of the cylinder was found to produce no change in the black-body efficiency, doubtless because the increase in emissivity of the interior was being compensated by an increase in diffuse as opposed to specular reflection. Covering the inside of the specimen cylinder with various blacks and oxide coatings was observed also to be ineffective. The provision of hot ends for the cylinder was seen to be impracticable on account of the difficulty of heating these ends separately, and further, because the low heat capacity of the specimen and its high axial temperature-gradient, would have made the formation of a uniform-temperature enclosure almost impossible.

It was also shown that the angle of view and aperture of the observing instrument did not, within reasonable limits, affect the black-body efficiency. It was, however, found that if the cylinder was dented near the electrodes the efficiency of the black-body was increased

This seemed to indicate that the finite length of the cylinder was responsible for the error since some radiation was being lost at the cold ends of the cylinder. A qualitative study of this may be made by supposing only a finite number of inter-reflections (instead of an infinite number) to take place inside the cylinder. In the case of four inter-reflections, for instance, the radiation from the " black-body " slit is given by

$$EB(1+R+R^2+R^3+R^4)=\frac{EB}{1-R}(1-R^5)=(1-R^5)B,$$

showing a black-body deficiency of R^5. The effect of this on the apparent emissivity may be seen from the following figures for platinum :—

	True E	Calculated deficiency	Apparent E (calc.)
0·65 μ	0·35	11%	0·39
4·0 μ	0·10	60%	0·25

These results are of the order of magnitude found and hence the hypothesis of a finite number of inter-reflections was felt justifiable. To minimize the effect, a special type of specimen was designed in which the middle portion of the cylinder was bulged out into a sphere. It was constructed of platinum by the Baker Platinum Co. The specimen was made in two halves, each consisting of a half cylinder bulged in the middle into a hemisphere. The wall thickness was such that when heated electrically the temperature of the spherical portion was practically uniform. A small slit 1 mm. wide and 1 cm. long on the edge of one hemisphere was used as a black-body source. The emissivity readings obtained with this spherical specimen showed a large improvement over those with the cylinder. The results, indeed, compared so favourably with the available data for platinum, that it was assumed, in default of any conclusive check, that the spherical black-body was sensibly perfect, at least in the near infra-red. It must, however, be pointed out that all absolute readings subsequently derived are liable to be high, the error increasing with wave-length. If these " sphere " results be taken as standard for platinum, then by comparing them with the data obtained for the same metal with the cylindrical specimen, a calibration curve may be drawn to correct " cylinder " values to the " sphere " standard. As an additional argument in favour of the adoption of this standard it may be noted that the results for iron and other metals obtained from the cylinder technique produce values comparing favourably with those of other workers when corrected in this manner. Moreover, extrapolation to the visible region indicates a correction to be applied to Ives' values. This correction is in good agreement with more recent determinations by other methods. Fuller details of the sphere/cylinder calibration are given in the following section.

§ 6. EXPERIMENTAL OBSERVATIONS

Platinum. All the specimens used were of 99·8 % purity platinum obtained from the Baker Platinum Co. In all, six full sets of observations were made with the cylindrical type of specimen and eleven with the spherical specimen. The agreement between different sets is of the order of accuracy of 1% for trend and 5% for absolute value. The average of each collection of data was taken. Smoothed values have been used in figure 2. It must be noted that since most of the possible errors tend to increase the observed emissivity, it is likely that a small constant error is introduced by taking the average of many sets of data, thus making all values slightly too high. The values of $E_{0·65}$ deduced from the pyrometer readings are estimated to be accurate to about 4%.

On a number of occasions the cylindrical specimen was photographed by its own radiation. The variation of emissivity with angle of emission was calculated from the intensity distribution across the cylinder. The results were not very accurate, but as a rough figure it may be stated that at a mean wave-length of 0·55 μ there exists a sharp maximum of about 120 % normal emissivity for an angle of emission of about 85°. (Stephens finds a maximum of 124% at 80° for platinum under similar conditions.)

Using the two sets of values obtained for the emissivity of platinum, a correction curve was plotted showing the calculated efficiency of the cylindrical type of black-body (that is, assuming the spherical type to be perfect). Since the two techniques had been used for platinum under approximately the same conditions and over slightly different ranges and intervals it was thought worth while to synthesize the seventeen sets of results. The final figures (smoothed) for the emissivity of platinum at an average temperature of 1 125° are as follows:—

Table 1

λ_μ	0·65	1·0	1·1	1·2	1·3	1·4	1·5	1·75	2·0	2·25
E (corr.)	0·330	0·293	0·287	0·284	0·284	0·276	0·270	0·255	0·240	0·228
λ_μ	2·5	2·75	3·0	3·25	3·5	3·75	4·0	4·25	4·5	4·75
E (corr.)	0·218	0·206	0·196	0·188	0·180	0·172	0·165	0·157	0·150	0·145

Because the correction curve must be independent of the nature of the emitting metal and can be a function only of its true emissivity and reflectivity, it is possible to use this curve to correct readings obtained for other metals with the "cylinder" technique. Before this can be done, however, two points must be considered:

(*a*) The curve must be extrapolated at both ends to cover the range of readings found for the other metals to be investigated. Bearing in mind the qualitative theory already given of this type of deficiency, and also the work of Buckley on similar problems, a reasonable extension of the curve was made. This is shown in figure 2, which gives the observed values for platinum together with full curves showing the black-body efficiency (B) plotted against (i) the true emissivity (E sphere) and (ii) the apparent emissivity (E cylinder). The use of this latter curve enabled all experimental results to be corrected by multiplying by the appropriate value of B.

(*b*) It will be seen that apart from the full curve shown, the experimental values indicate a secondary component in the correction curve. This subsidiary deficiency has a "cocked hat" form with a maximum of the order of 4%. In view of the fact that this maximum coincides with the maximum in the radiant energy curves for the temperatures used, it is probably due to the scattering of radiation or to a heating effect involving re-radiation from the spectrometer slits, etc. Assuming some explanation of this nature, a correction may easily be made by plotting the extra deficiency (C) against wave-length for platinum and applying this to the results for other metals. A graph of this correction is inset in figure 2. In spite of the somewhat arbitrary nature of this second correction it was thought sufficient since the effect of it is only of the same order as the experimental error. In all subsequent results, values have been corrected by multiplying by the appropriate value of B and then subtracting a percentage C.

Iron. The metal used in all experiments was obtained in the form of 5 mm. diameter electrode rods produced for spectroscopic work by Messrs. Johnson, Matthey. An analysis indicated a purity of 99·96%. The metal was rolled into thin sheet between clean steel rollers. Throughout the observations no deterioration of the surface was observed even after heating in a hydrogen atmosphere for over 12 hours at 1200° c. In all, four different specimens were used and, with these, nine sets of observations were obtained. The observed values were corrected and smoothed graphically to give final figures, the accuracy of which is estimated to be similar to that for platinum. These figures for pure iron at an average temperature of 1245° c. are as follows:—

Table 2

λ_μ	0·65	1·0	1·1	1·2	1·3	1·4	1·5	1·75	2·0
E (cyl.)	0·494	0·397	0·389	0·378	0·368	0·363	0·357	0·341	0·334
E (corr.)	0·437	0·340	0·330	0·316	0·306	0·298	0·290	0·270	0·260

λ_μ	2·25	2·5	2·75	3·0	3·25	3·5	4·0	4·5
E (cyl.)	0·326	0·323	0·315	0·312	0·306	0·301	0·288	0·282
E (corr.)	0·252	0·248	0·244	0·240	0·237	0·235	0·225	0·218

Molybdenum. High purity electrode rods of this metal proved too brittle for rolling into sheet in the same manner as iron. Thin sheet of ordinary chemical purity, obtained from Messrs. Johnson, Matthey, was therefore used. An excellent polished surface free from blemishes was easily obtained and maintained throughout the experiments. The metal was heated in a static atmosphere of hydrogen and 16 sets of observations were obtained. These were averaged and corrected from the calibration curves. The mean temperature of the specimens was 1226° c. The average accuracy of each observed point is estimated at 2 %.

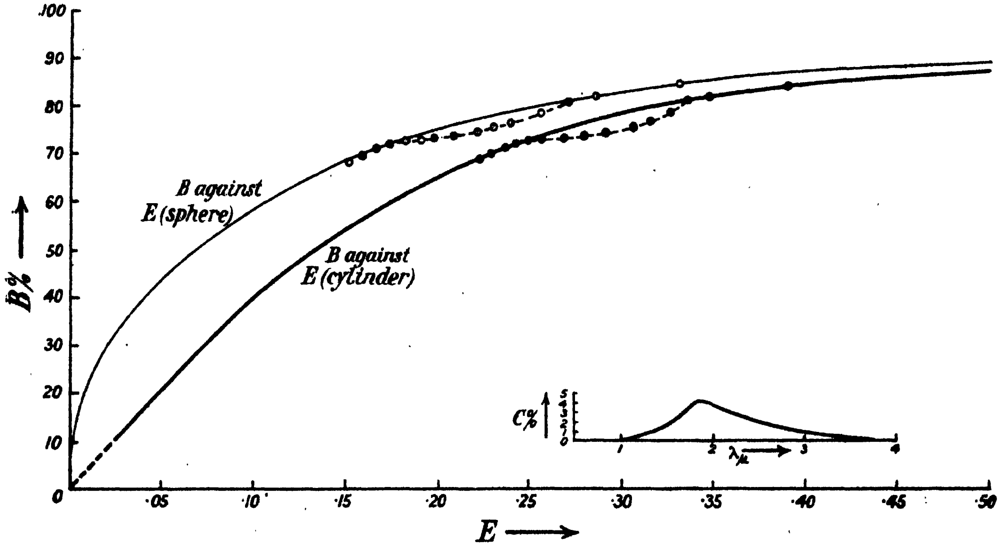


Figure 2. Efficiency of black-body.

Table 3

λ_μ	0·65	1·0	1·1	1·2	1·3	1·4	1·5	1·75	2·0
E (cyl.)	0·360	0·410	0·387	0·372	0·351	0·335	0·319	0·297	0·269
E (corr.)	0·299	0·350	0·327	0·308	0·288	0·270	0·251	0·220	0·197

λ_μ	2·25	2·5	2·75	3·0	3·25	3·5	3·75	4·0	4·5
E (cyl.)	0·264	0·243	0·242	0·223	0·214	0·208	0·200	0·193	0·165
E (corr.)	0·185	0·170	0·163	0·151	0·142	0·135	0·129	0·122	0·095?

Copper. High grade copper of "electrical" purity was used in all observations. Some difficulty was experienced in working with this metal, both on account of its low melting point and also because of its sudden fusion, caused no doubt by recrystallization taking place at high temperatures. Readings had to be restricted to the range 1·5 to 4·0μ on account of the small quantity of energy emitted at the low working temperature (901° c.). Ten sets of observations were obtained and averaged. The surface appeared bright and uncontaminated by impurity, but in spite of this the average error was somewhat high (5–10%). Another source of error is the extrapolation needed in the correction curve to permit its use in connection with the low emissivity of copper, and for these reasons the results for this metal may be rather unsatisfactory.

Table 4

λ_μ	0·65	1·5	2·0	2·5	3·0	3·5	4·0
E (cyl.)	0·420	0·150	0·141	0·118	0·105	0·102	0·088
E (corr.)	0·360	0·079	0·065	0·052	0·043	0·038	0·032

Nickel. As with iron, the metal was obtained from Messrs. Johnson, Matthey in the form of rods produced for spectroscopic purposes and estimated to contain 99·97% nickel. The rod was rolled into thin sheet between clean steel rollers. No trouble was found in producing and maintaining a bright, clean surface, although buckling of the specimen was more pronounced with this metal than with any other used. Because of this a number of the sets had to be abandoned and discarded. In spite of this twelve complete runs of observations were made and the results averaged. These readings were corrected from the curve. The emissivity of pure nickel at an average temperature of 1110° c. was found to be:—

Table 5

λ_μ	0·65	1·0	1·1	1·2	1·3	1·4	1·5
E (cyl.)	0·600	0·358	0·373	0·358	0·345	0·336	0·319
E (corr.)	0·540	0·292	0·314	0·292	0·280	0·270	0·250

λ_μ	2·0	2·5	3·0	3·5	4·0
E (cyl.)	0·296	0·279	0·256	0·241	0·231
E (corr.)	0·220	0·205	0·187	0·174	0·162

§7. CRITICAL SUMMARY OF EXPERIMENTAL OBSERVATIONS

For convenience the final, corrected and smoothed values for the high-temperature emissivities of the five metals investigated have been tabulated together. The figures enclosed in brackets are doubtful, but apart from these a mean deviation from the smooth curve of the order of 0·5% for the emissivity values can be claimed (with the reservations made in the section on black-body efficiency).

Table 6. Summary of data (emissivities, %)

λ_μ	Pt 1125° c.	Fe 1245° c.	Cu 901° c.	Mo 1226° c.	Ni 1110° c.
0·65	22·0	43·7			(54·0)
1·0	29·3	34·0		35·0	(29·2)
1·1	28·7	33·0		32·7	31·4
1·2	28·4	31·6		30·8	29·2
1·3	28·4	30·6		28·8	28·0
1·4	27·6	29·8		27·0	27·0
1·5	27·0	29·0	7·9	25·1	25·0
1·75	25·5	27·0		22·0	
2·0	24·0	26·0	6·5	19·7	22·0
2·25	22·8	25·2		18·5	
2·5	21·8	24·8	5·2	17·0	20·5
2·75	20·6	24·4		16·3	
3·0	19·6	24·0	4·3	15·1	18·7
3·25	18·8	23·7		14·2	
3·5	18·0	23·5	3·8	13·5	17·4
3·75	17·2			12·9	
4·0	16·5	22·5	3·2	12·2	16·2
4·25	15·7				
4·5	15·0	21·8		(9·5 ?)	
4·75	14·5				

From this table we find:—

Platinum. The anomaly giving a minimum at 1·1 μ and a maximum at 1·2 μ occurred in all series taken in this region and, therefore, is unlikely to be spurious. In support of this it may be noted that McCauley (1913) found a similar maximum at 1·1 μ for platinum at 1300° c. The size of the anomaly found in the present work is less than that noted by McCauley but against this is the fact that according to his work the maximum is reduced at 1000° c. to a point of inflexion. Although the present work is in good agreement with McCauley for $\lambda > 3\,\mu$, it becomes increasingly low at short wave-lengths and eventually approximates in the visible region to the room-temperature data of Hagen and Rubens (1903) and Coblentz (1911). Incidentally, the data of these workers and of Försterling and Freedericksz (1913) all show fluctuations from a smooth curve in the interval 0·9 to 1·5 μ, although in no case have sufficient points been taken to show that this is not an example of random error. The value of $E_{0\cdot65}$ is in excellent agreement with that given by Foote, Fairchild and Harrison (1921). At the other end of the wave-length range a comparison may be had by extrapolating the data given by Hagen and Rubens (1909) for the emissivity at 6·65 μ (using a *Reststrahlen* method). In the range 100° to 500° c. they found $E = 0\cdot0383\ (1 + 0\cdot00162T)$. This formula gives $E = 0\cdot108$ for 1125° c., this point falling on the present curve. Since the curve lies wholly above the corresponding room-temperature values, it will be seen that there can exist no real X-point in the range covered. This conflicts with experiments carried out by Hagen and Rubens (1910) at 2, 4, and 6 μ between 635° and 1455° c. Their results, which are, unfortunately, on an arbitrary scale, show that at 4 μ and 6 μ the Hagen–Rubens relation is valid, but for 2 μ a negative temperature coefficient occurs indicating an X-point at about 2·15 μ. On the basis of the present work and by the comparison which has been made with other data for short wave-lengths, it is believed that these measurements of Hagen and Rubens are erroneous.

Deficiency correction. An interesting application of the deficiency curve found experimentally for a cylinder "black-body" is the correction of Ives's (1924) standard of visible radiation. Ives found the brightness of his standard (of the same type and size as the present specimen) to be 55·40 candles/cm^2 at the freezing point of platinum. For the visible region the mean emissivity of platinum at 1773° c. is approximately 0·35, for which the extrapolated calibration curve indicates a "black-body" efficiency of about 86 %. This increases the figure given by Ives to 64·5 cp./cm^2, which is in better agreement with the more modern determination of 61·0 cp./cm^2 Moreover, Ives reported that the colour temperature of his source was unaccountably 35° c. too high. A graphical computation involving the visibility function and estimated "black-body" deficiencies at different wave-lengths showed that such a figure as 30–40° error in the colour temperature should reasonably occur.

Iron. The observed curve was free from any anomalies. No previous measurements have been made on the pure metal in this region of the spectrum at high or low temperatures. Indeed, the only high-temperature results existing for steel are due to Hagen and Rubens (1910). They cover, roughly speaking, the range 0–300° c. and 0·78–5·0 μ, 6·65 μ, 8·85 μ and 25·5 μ. Since the steel used had a very high nickel content it is futile to attempt any quantitative comparison of results. The available figures for reflectivity (at room temperature) are two sets by Hagen and Rubens ("ungehärtet Stahl"), one by Coblentz (iron of unspecified purity) and one by Ingersoll (1910) (tool steel). All these curves intersect the present high-temperature curve at points ranging from 1·0 to 1·5 μ. This is in qualitative agreement with the high-temperature observations of Hagen and Rubens which indicate an X-point in the neighbourhood of 1·0 μ.

Molybdenum. The only previous data existing for the infra-red emissivity of this metal are those derived from the room-temperature reflectivity figures given by Coblentz (1911), and the only high-temperature values are those due to Whitney (0·67 μ) and to Worthing (0·47 μ) and (0·67 μ). The present curve (which shows no anomalies) has the same trend as that given by Coblentz but a point of intersection occurs in the region of 1·4 μ. Below this point the temperature coefficient is negative and above it positive. The existence of such a negative coefficient is in agreement with Worthing's (1925) investigation which showed that at 0·47 μ the emissivity decreased from about 0·42 at 400° K. to 0·36 at 2800° K. Similarly at 0·67 μ the decrease was from 0·42 at 400° K. to 0·33 at 2800° K. This is, however, at variance with the finding of Whitney (1935) who reported that between the temperatures of 1400° K. and 2000° K. the emissivity at 0·67 μ was constant. It is to be suspected that the small variation was masked by experimental error or else that the properties of a specimen subjected to heat-treatment and "baking-out" are different from those of the metal used, both in the present work and in that of Worthing. The present value for $E_{0\cdot65}$ is in good agreement with Worthing's and fits well on the extended curve for the infra-red region.

Copper. As has been previously noted, the low emissivity of this metal considerably reduces the accuracy of the present method. In spite of this the values have the same trend and order of magnitude as those found by Hurst (1933), to whom the only previous infra-red measurements at high temperatures are due. The present curve intersects that of Hurst in the region of $2{\cdot}9\,\mu$. At $1{\cdot}5\,\mu$ it is 2% higher and at $4\,\mu$ it is about 1% lower than Hurst's values. Infra-red measurements at room temperature have been made by Försterling and Freedericksz (1913), by Hagen and Rubens (1903) and by Ingersoll (1910). The present curve lies well above all these, indicating the absence of any X-point in this region. This is in agreement with Bidwell (1913) who finds a small positive temperature coefficient for both the solid and liquid metal at $0{\cdot}6\,\mu$, but in disagreement with Burgess (1909) who measured a small negative coefficient for the liquid metal in this wave-length region. Schubert (1937) has also investigated the reflectivity of the hot metals but his finding is that no true coefficient could be detected within the experimental limits.

Nickel. The present curve is uniformly 5 % higher than the figures found by Hurst under similar conditions. It lies wholly above the room-temperature curve of Hagen and Rubens (1902–3) and hence no X-point is found. It has already been noted that buckling of the specimen was rather troublesome in the case of nickel, and this may well have resulted in a larger black-body deficiency and hence to the 5 % discrepancy with Hurst's results. This is made more likely since a number of workers have noted an X-point for this metal (e.g. Hurst $\lambda_x = 1{\cdot}85\,\mu$, Cennamo (1939) $\lambda_x = 2{\cdot}25\,\mu$, Rubens (1910) $\lambda_x = 1{\cdot}4\,\mu$, Reid $\lambda_x = 2{\cdot}1\,\mu$). This is supported by the work of Bidwell, who finds a negative temperature coefficient in the visible, but in opposition Wahlin and Wright (1942) find that with a properly baked-out specimen the emissivity is sensibly constant (cf. Whitney's result for copper). The value obtained for $E_{0{\cdot}65}$ is a reasonable continuation of the present curve, but again this result is too high to give a temperature coefficient of the order of that obtained by other workers. It must therefore be presumed that this curve is about 5 % too high and that consequently an X-point may occur in the region of $2{\cdot}0\,\mu$.

ACKNOWLEDGMENTS

This research forms part of a programme being undertaken for the Foundry Steel Temperature Sub-Committee of the Steel Castings Research Committee of the British Iron and Steel Research Association and Iron and Steel Institute, with their financial support.

In conclusion I wish to thank Dr. H. Lowery, Principal of the S.W. Essex Technical College, for his direction of the research.

REFERENCES

ANGELL, 1911. *Phys. Rev.*, **33**, 422.
BEEKMAN and OUDT, 1925. *Z. Phys.*, **33**, 831.
BIDWELL, 1913. *Phys. Rev.*, **1**, 482; 1915. *Ibid.*, **3**, 439.
BUCKLEY, 1927. *Phil. Mag.*, **4**, 753; 1928. *Ibid.*, **6**, 447; 1934. *Ibid.*, **17**, 576.
BURGESS, 1909. *Bull. Bur. Stand.*, **6**, 111.
CENNAMO, 1939. *Nuovo Cimento*, **16**, 253.
COBLENTZ, 1908. *Bull. Bur. Stand.*, **5**, 339; 1911. *Ibid.*, **7**, 197; 1918. *Ibid.*, **5**, 312.

Cunnold and Milford, 1934. *Phil. Mag.*, **18**, 561.
Foote, Fairchild and Harrison, 1921. *Sci. Pap. Bur. Stand.*, 170.
Försterling and Freedericksz, 1913. *Ann. Phys., Lpz.*, **40**, 201.
Hagen and Rubens, 1900. *Ann. Phys., Lpz.*, **1**, 352; 1902 a. *Ibid.*, **8**, 1; 1902 b. *Ibid.*, **8**, 432; 1903. *Ibid.*, **11**, 873; 1909. *Abh. Preuss. Akad. Wiss.*, **16**, 478; 1910. *Ibid.*, **23**, 467.
Hurst, 1933. *Proc. Roy. Soc.*, A, **142**, 466.
Ingersoll, 1910. *Astrophys. J.*, **32**, 265.
Ives, 1924. *J. Franklin Inst.*, **197**, 147.
Langmuir, 1916. *Phys. Rev.*, **7**, 302.
McCauley, 1913. *Astrophys. J.*, **37**, 164.
Mendenhall, 1911. *Astrophys. J.*, **33**, 91.
Price and Lowery, 1944. *J. Iron and Steel Inst.* (Steel Castings Research Committee), **149**, 523.
Reid, 1941. *Phys. Rev.*, **59**, 161.
Rubens, 1910. *Phys. Z.*, **11**, 139.
Schubert, 1937. *Ann. Phys., Lpz.*, **29**, 473.
Stephens, 1936. *J. Opt. Soc.*, **29**, 158.
Wahlin and Wright, 1942. *J. Appl. Phys.*, **13**, 40.
Whitney, 1935. *Phys. Rev.*, **48**, 458.
Worthing, 1925. *Phys. Rev.*, **25**, 846; 1926. *Ibid.*, **28**, 174.
Worthing, 1940. *J. Appl. Phys.*, **11**, 421.

THE TEMPERATURE VARIATION OF THE EMISSIVITY OF METALS IN THE NEAR INFRA-RED

By DEREK J. PRICE,
S.W. Essex Technical College, London

Communicated by H. Lowery; MS. received 15 *July* 1946

ABSTRACT. Evidence for the existence of a wave-length, peculiar to each metal, for which the temperature coefficient of emissivity is zero is analysed and commented upon. Further experimental data obtained from the author's measurements of infra-red emissivity of hot metals are submitted in support, and it is shown that the appearance of such an *X-point* phenomenon would explain many of the divergent results obtained for the temperature variation of emissivity in the visible region, besides providing a new series of characteristic wave-lengths for metals.

The Hagen and Rubens (1903) approximation has been tested experimentally for a number of metals and has been shown to account accurately for the emissivity and the temperature coefficient of emissivity in all cases for wave-lengths greater than about 10μ. For shorter wave-lengths, experimental results diverge from this theory until, in the visible region, the Hagen–Rubens approximation is found to yield no qualitative picture of either emissivity or its temperature variation. The position is further obscured by the fact that while the experimentally observed temperature coefficient of emissivity in the visible is always much smaller than that demanded by this formula, the data show large discrepancies between different workers. As an example may be cited a collection of data given by Worthing (1941) for the case of platinum (figure 1). The effect sought is certainly much smaller than that expected from the Hagen–Rubens formula, indeed it is to be presumed that just because it is so minute, the existence of any small methodical error or a slight change of surface could easily produce just the discrepancy noted in these results. The appearance of conflicting data is not confined to the case of platinum, but is to be found in nearly all investigations of temperature variation of the emissivity, reflectivity, and optical constants of metals in the visible region.

A most interesting effect, contrary to all established notions, has been found by many workers who have observed a *decrease* in emissivity with rise in temperature. The existence of such a negative temperature coefficient in the visible region must imply that somewhere between it and the region of validity of the Hagen–Rubens relationship (where the temperature coefficient must be positive) there is a wave-length where the emissivity is constant with temperature, that is, the coefficient is zero. This point, here for shortness termed the *X-point*, has been observed directly in at least four investigations made on the temperature variation of reflectivity of metal surfaces. (Since, by Kirchhoff's Law, the sum of the reflectivity and emissivity of a non-transparent medium must be unity, these results may be used to calculate emissivity values). The data given by these four investigations are by no means in complete accordance. For instance, Rubens (1910) gives an X-point for nickel at 1·25 μ whereas Reid (1941) finds it at 2·15 μ, and in the case of tungsten, Weniger and Pfund (1919) note the X-point at 1·27 μ and Ornstein (1936) (collating the observations of Hamaker, Vermeulen and van Zelst) finds it at 1·64 μ. The effect is so small, and the disturbance produced by heating a polished metal surface so large, that the exact location of an X-point by these direct means is seen to be a matter of great difficulty, and it is therefore hardly surprising to find the discrepancies noted.

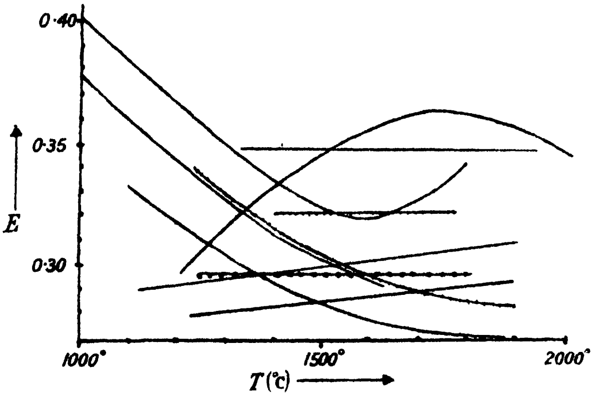


Figure 1. $E_{0\cdot6\,\mu}$ for platinum (adapted from Worthing, 1941, p. 1180).

Both Rubens (for nickel) and Ornstein (for tungsten) find the emissivity to increase linearly with temperature. The data of the latter worker are very striking, for when they are plotted, it can be seen that all the curves of emissivity against wave-length at temperatures ranging from 300° K. to 3000° K. intersect accurately at one point. Unfortunately it is not clear from the original papers whether measurements in the region of the X-point were actually made at the large number of temperatures quoted, or whether these figures have been interpolated from readings taken at the two limiting temperatures only. Certainly an X-point does exist here, and, if the results have not been obtained by interpolation, it would appear that this wave-length is a constant property of the metal over an unusually great range of temperature. In any case, the existence of a specific wave-length in the near infra-red region, characteristic of each metal, is sufficiently interesting to merit further examination, even if this X-point is not absolutely constant at all temperatures.

Since the effect sought appears to be so small as to be almost masked by the experimental errors associated with its direct observation, it is necessary to resort to an indirect attack . Data found for high temperatures by an emissivity technique may be compared with data found for room temperatures from the measurement of reflectivity, and from this information the temperature coefficient may be calculated and the X-point (if any) detected and located. Two main criticisms may be made of this method:—

1. It is necessary to assume a linear variation of emissivity with temperature in order to calculate the coefficient. Since, however, the effect is small and also any deviation from a linear relation would not influence the location of an X-point, this criticism may be neglected.
2. Different experimental techniques and even different specimens have been used for the two temperatures. Against this it may be said that only figures obtained for the pure metal surface should be used, and whenever possible the data compared should be the mean of a number of independent investigations.

In the case of nearly all the common metals there exist sufficient data for room temperatures, e.g. those obtained in the classical investigations of Coblentz and of Hagen and Rubens, and hence wherever experiments have been made at high temperatures it is possible to carry out the above comparison.

Leaving for the moment a number of points of intersection occurring in the visible region, the following X-points obtained by this method may be added to the four previously noted:—

Author for high-temperature results	Metal	X-point (μ)
Hurst, 1933	Cu	2·2
Hagen and Rubens, 1910	Fe	1·0
Hurst, 1933	Ni	1·8
Cennamo, 1939	Ni	2·5
McCauley, 1913	Pd	1·0
Davisson and Weeks, 1924	W	1·28
Forsythe and Worthing, 1925	W	1·3

Again, it is found that the agreement is not very good, due doubtless to the partial masking of the effect sought by the experimental and systematic errors. In the transition region between this X-point and the limit of validity of the Hagen–Rubens relationship, the temperature coefficient is larger and is no longer masked by the experimental inaccuracies. Data obtained by the present author (1947) for the infra-red emissivity at high temperatures of a number of metals enables the comparison method described above to be used for determining the variation of the temperature coefficient of emissivity with wave-length. The room-temperature values employed in this comparison have all been obtained from a smooth curve drawn through the mean of all available results for each metal in the ultra-violet, visible, and infra-red regions.

Although a complete examination of the temperature coefficient, obtained for each wave-length by measuring the emissivity at a number of temperatures, would be more desirable than the room-temperature and single high-temperature here employed, this plan was found impossible. The available range (*circa* 900 °c. to the melting point) was too small, and the systematic errors inherent in the black-body calibration and comparison were too large to enable useful or consistent figures to be obtained for the coefficient, using an emissivity technique of this nature.

The existence of a complete curve for the variation of temperature coefficient with wave-length between the Hagen–Rubens limit and the X-point region, makes possible for the first time a fairly accurate location of this X-point to be made by extrapolation of the smooth curve obtained to cut the wave-length axis at the point of zero temperature coefficient. As will be seen, in some cases, the experimental curve departs in the X-point region from the trend found for longer wave-lengths. This may be either a real phenomenon due to the proximity of the absorption bands of the bound electrons, or it may in some measure be due to the experimental and systematic errors previously noted.

Let us examine in more detail the nature and trend of these temperature coefficient curves (figure 2*a*). Starting at the long wave-length end, we see that above the limit of validity of the Hagen–Rubens relationship the temperature coefficient of emissivity must be half the temperature coefficient of electrical resistivity of the metal. This follows from the fact that in this region emissivity is proportional to the square root of resistivity. As has been stated, the validity of this law has been conclusively established in the now classical work of Hagen and Rubens. Below the limit of validity, the temperature coefficient decreases almost linearly with wave-length up to the region of the X-point. At either end of the linear region the curve turns, at the upper end to meet the horizontal limit predicted by the Hagen–Rubens theory, and at the lower to come eventually to the origin. This latter point follows from the fact that the emissivity of all bodies must be unity at the limit of zero wave-length, and hence the temperature coefficient must be zero.

The temperature coefficient graphs deduced, as stated, from the present author's measurements on molybdenum, iron and nickel are shown in figures 2*c*, 2*d* and 2*e*. That for platinum has been given elsewhere (Price, 1946). The data for copper were somewhat inaccurate, due to the low temperature at which results had to be taken, and the only conclusion which could be drawn was that within the experimental error the temperature coefficient was of the order predicted by the Hagen–Rubens relationship for wave-lengths beyond 1·5 μ, i.e. the horizontal portion of the curve extended to much shorter wave-lengths than for any of the other metals investigated. The figures for molybdenum are very striking since they show clearly all the phenomena noted, an X-point at 1·8 μ, the turning round to the Hagen–Rubens limit starting point at about 2·5–3·0 μ, and the appearance of a minimum at 1·0 μ where the curve turns to go through the origin Iron has an X-point at 1·55 μ, but the turning of the curve at either end is only just indicated in the range covered by observation. The curve for nickel is again too short to show the turning, but besides this, the rather surprising fact is noticed that the temperature coefficient is positive throughout. This omission of the X-point noted by Rubens, Reid, Cennamo, and Hurst, is possibly due to a discontinuity at the Curie point vitiating the assumption of a linear increase with temperature from 0° c. to 1100° c. In this case an investigation over a range of temperatures using a more suitable technique (e.g. reflectivity) would be very desirable.

The existence of an X-point, whether real or virtual, is seen to provide a key both to the short-wave region and to the transition period. In the former case, between the X-point and the origin, the temperature coefficient is small and the curve may show considerable fine structure in the region of absorption bands. Data for tungsten calculated from Ornstein's values (figure 2*b*) indicate the sort of fine structure to be found. In some cases it is possible that one or more of these small maxima or minima should intersect the wave-length axis, so giving one or more further points of zero temperature coefficient. These "spurious X-points" occurring in the region of a band of selective reflection have been observed by Fujioka and Wada (1934) (Ag, 0·31 μ, Au, 0·47 μ, Cu, 0·5 μ), by Worthing (1921) (Au, 0·47 μ), by Ebeling (1925) (Cu, 0·5 μ), and by Ornstein and Van der Veen (1939) (Fe, 0·45 μ ?). It is, of course, somewhat difficult to distinguish between the real and the spurious X-points. Indeed, the position may be said to be satisfactory only when a complete curve of temperature coefficient against wave-length is known. In the cases quoted, however, the presence of a known absorption band at the stated wave-length makes it evident that the phenomenon is due to the temperature variation of this band. This type of explanation cannot serve for the X-point values previously quoted, since there seem to be no known strong resonance frequencies for metals in the region 1·0–2·5 μ.

The only comparative account of this X-point phenomenon is due to Worthing (1926) who, in a short paragraph, quotes his own result for gold ($\lambda_x = 0{\cdot}46$) and that of Weniger and Pfund (1919) for tungsten ($\lambda_x = 1{\cdot}27$), and from these values conjectures that the X-point wave-lengths may be proportional to the melting points of the elements (Au = 1336° K., W = 3655° K.). This is at variance with the fact that λ_x for nickel is of the same order as that for tungsten and in addition it is felt that the divergence of the values given by different workers is too high to allow any law to be formulated on the basis of just two random results. Further it is possible that the figure quoted for gold represents a "spurious" X-point due to the influence of the absorption band in the visible region.

There are at least two possible explanations of the X-point phenomenon, but in neither case is the associated theory sufficiently developed to provide a more exact picture. It is well known that, in general, an absorption band shifts towards higher wave-lengths and becomes broader with increase in temperature. It has also been noted by Schaum and Wustenfield (1911) that the broadening is more pronounced on the long wave side of the band. This type of change is shown in figure 3, and is seen to entail an intersection of the two curves, i.e. an X-point.

It is possible that a similar type of explanation should be applicable to the free-electron portion of the metal. This would mean that the half-width of the

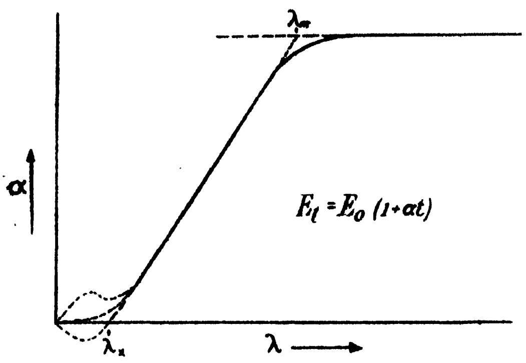


Figure 2 *a*. Idealized curve of temperature coefficient of emissivity against wave-length (cf. Price, 1946).

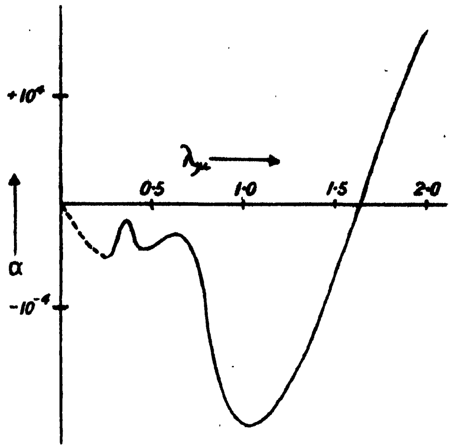


Figure 2 *b*. Temperature coefficient for tungsten at 1600° K. (after Ornstein, 1936).

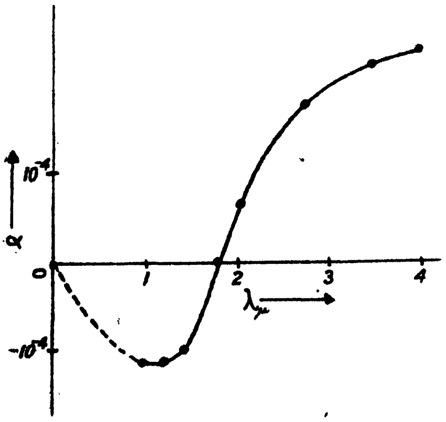


Figure 2 *c*. Temperature coefficient for molybdenum.

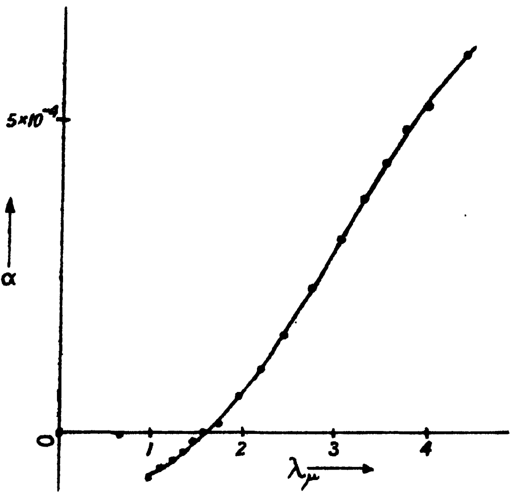


Figure 2 *d*. Temperature coefficient for iron.

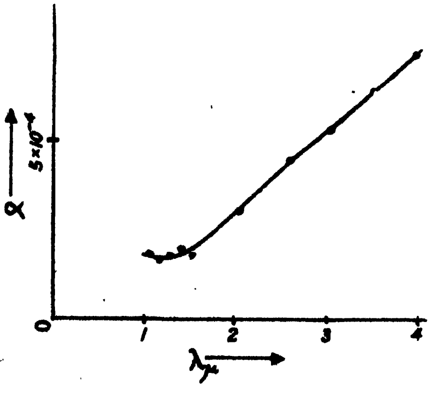


Figure 2 *e*. Temperature coefficient for nickel.

absorption band of the free electrons should decrease with increase of temperature while the long wave-length values of emissivity should, in accordance with the Hagen–Rubens relationship, increase with temperature. This type of explanation is illustrated diagrammatically in figure 4 (*a*). The objection to this explanation is that usually the width of an absorption band increases rather than decreases with rise in temperature.

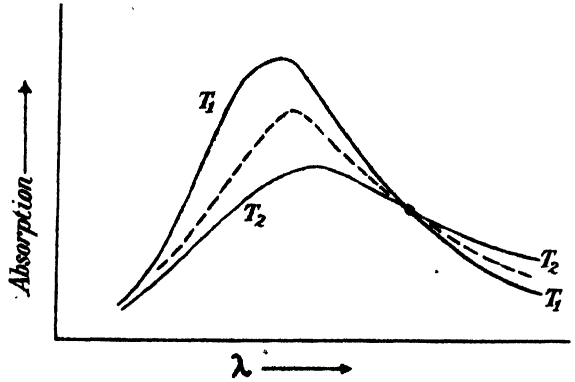


Figure 3. Influence of rise of temperature on an absorption band.

An alternative explanation is that the X-point represents a balance between two opposing effects due on the one hand to the free-electron contribution, and on the other to a bound-electron absorption band. This type of effect is shown diagrammatically in figure 4 (*b*), where it may be seen that the rise in emissivity due to the free-electron portion is balanced at the X-point by the broadening of the band due to one class of bound electrons. It is probably this effect which produces

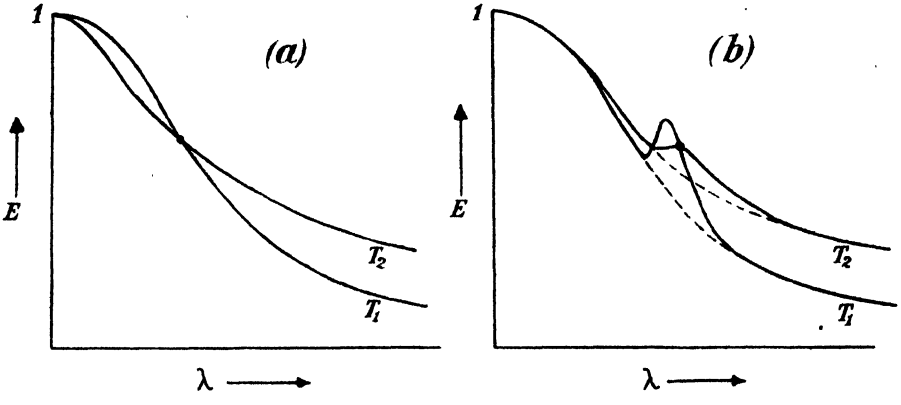


Figure 4. Influence of rise of temperature on emissivity curves.

the "spurious" X-points previously noted in the region of the absorption bands at 0·25–0·6 μ. As has been said, however, the absence of strong resonance wave-lengths in the region 1·0–2·5 μ makes this explanation untenable in the case of the true X-point. Now although no selective effects have been found at room temperature, a few investigators have noted that at high temperatures a new absorption band appears in the near infra-red in the very region demanded by the X-point. For instance, Ornstein's data (1936) for tungsten show a new maximum developing at about 2·0μ, and McCauley's investigations (1919) of tantalum, platinum and palladium all show new maxima appearing at high temperatures in the region of 1·0μ. This result for platinum has been confirmed by the present author, who detected a maximum at 1·25μ at 1125° c. In all these cases the maximum and minimum appear to become more pronounced with increase in temperature, and it is quite possible that the appearance of the selective absorption region, hidden at room temperatures, is responsible for the X-point phenomenon. Little more can be said of this possibility without a full catoptric investigation of the optical constants of metals at high temperatures.

Certainly it can be seen that emissivity methods are not sensitive enough to measure accurately the small temperature coefficients existing near the X-point and in the visible region, although some information has been obtained from a series of measurements at longer wave-lengths. This information, in the form of temperature-coefficient graphs, shows clearly the manner in which the Hagen–Rubens relationship breaks down at short wave-lengths. Further, these graphs are characterized by the appearance of an X-point which not only entails the existence of negative temperature coefficients, but also implies that an explanation of the divergent results obtained by many workers may be found if the sensitivity of this X-point to surface contamination, crystal structure, etc. is assumed.

ACKNOWLEDGMENT

I wish to thank Dr. H. Lowery, Principal of the S.W. Essex Technical College, for his direction of the research.

REFERENCES

Cennamo, 1939. *N. Cimento*, **16**, 253.
Davisson and Weeks, 1924. *J. Opt. Soc. Amer.*, **8**, 581.
Ebeling, 1925. *Z. Phys.*, **32/7**, 489.
Forsythe and Worthing, 1925. *Astrophys. J.*, **61**, 146.
Fujioka and Wada, 1934. *Sci. Pap. Inst. Phys. Chem. Res., Tokio*, **25**, 9.
Hagen and Rubens, 1903. *Ann. Phys., Lpz.*, **11**, 873.
Hagen and Rubens, 1910. *Abh. Preuss. Akad Wiss.*, **23**, 467.
Hurst, 1933. *Proc. Roy. Soc.*, A, **142**, 466.
McCauley, 1913. *Astrophys. J.*, **37**, 164.
Ornstein, 1936. *Physica*, **3**, 561.
Ornstein and van der Veen, 1939. *Physica*, **6**, 439.
Price, 1946. *Nature. Lond.*, **157**, 765.
Price, 1947. *Proc. Phys. Soc.*, **59**, 118.
Price and Lowery, 1944. *J. Iron Steel Inst.*, no. 1, 523.
Reid, 1941. *Phys. Rev.*, **59**, 161.
Rubens, 1910. *Phys. Z.*, **11**, 139.
Schaum and Wustenfield, 1911. *Z. Wiss. Photogr.*, **10**, 213.
Weniger and Pfund, 1919. *Phys. Rev.*, **14**, 427.
Worthing, 1921. *J. Franklin Inst.*, **192**, 112.
Worthing, 1926. *Phys. Rev.*, **28**, 174.
Worthing, 1941. *Temperature, Its Measurement and Control in Science and Industry* (American Institute of Physics; Reinhold Publishing Corp.), p. 1164.

THE FREEZING-IN OF NUCLEAR EQUILIBRIUM

By A. R. UBBELOHDE,
Queen's University, Belfast

MS. received 25 *October* 1946

ABSTRACT. Experimental values of abundance ratios of stable isotopes, and of their mass differences, are used in a graphical method of testing how far the reaction velocities for different nuclear transformations occur at comparable rates under the conditions of freezing-in. With the limited data so far available for exact mass differences (up to atomic mass 57, Fe), a single freezing-in parameter of approximately 10^{10} degrees centigrade is found to account for a large proportion of the stable isotopes. However, a different origin, or very different reaction velocities for the nuclear transformations, is indicated for some of the stable isotopes, such as those of lithium.

§1. INTRODUCTION

Many examples are known of the freezing-in of molecular or chemical equilibrium, and some of the features of such frozen equilibria may be transposed to the problem of the freezing-in of nuclear equilibria.

Freezing-in occurs as the temperature or density of a system in equilibrium is progressively lowered, when the velocity with which changes of concentration can occur eventually becomes insufficient to adjust these concentrations to the fall in temperature, or density. If the only change occurring were the fall in temperature, Le Chatelier's principle indicates that the composition would lag behind that for true equilibrium, in favour of the endothermic components of the system. But in the case of nuclear equilibrium, freezing-in may have been determined by both temperature and partial pressure changes, and the characterization of freezing-in conditions is more complex. In the discussion which follows, the treatment is simplified deliberately, since no more detail seems to be justified with the experimental information available at present.

Assuming that the various atomic nuclei in the earth's substance were at some stage in equilibrium, this would call for detailed balancing between the various processes of nuclear transformation, so long as equilibrium was maintained. Processes involving a change in atomic number would include those verified in the laboratory, such as

$$ {}^{m}_{n}A + {}^{1}_{1}p \rightarrow {}^{m+1}_{n+1}A \qquad \ldots\ldots(1)$$

where ${}^{m}_{n}A$ is an atom of mass number m and charge n, and ${}^{1}_{1}p$ is a proton. Processes involving a change of atomic mass without change of atomic number would have included transformations such as

$$ {}^{m}_{n}A + {}^{1}_{0}n \rightarrow {}^{m+1}_{n}A \qquad \ldots\ldots(2)$$

where ${}^{1}_{0}n$ is the neutron. But other processes not yet discovered in the laboratory may also have been involved, and one of the purposes of what follows is to examine in what cases such processes may have been operative, and thus to throw fresh light on nuclear transformations.

When the statistical thermodynamics controlling equilibrium in a system is known, the concentrations reached in frozen-in equilibrium may be used to calculate a nominal "freezing-in temperature". This will have a value lying somewhere in the region where the rate of adjustment of concentrations is beginning to lag behind the rate of change of external conditions. In fact, such a freezing-in temperature may be regarded as a parameter characterizing the reaction velocities in the system under the conditions prevailing at the time.

Experimental data for the calculation of such parameters include (i) the relative abundance of elements of different atomic number, first discussed by Harkins (1917); and (ii) the relative abundance of isotopes of the same element. It does not necessarily follow that the freezing-in temperature for reactions in which the atomic number changes approximates to that for reactions in which only the atomic mass changes. Strictly speaking, each of the distinct processes involved in the detailed balancing which maintains equilibrium may freeze-in at a different temperature. It is therefore necessary to select one process at a time in comparing the freezing-in conditions for the different nuclei. Owing to the fact that chemical segregation of some of the elements subsequent to the freezing-in may have falsified some of the abundance ratios for nuclei of different atomic number, data of the type (i) will not be used in the present paper, though a similar procedure can be followed, as in the calculations given below, which are restricted to data of the type (ii).

Fairly accurate information has been collected about the abundance ratios of isotopes of the same element (Seaborg, 1944). These may be used to examine the hypothesis that equilibrium between isotopes of the same element was maintained at one period of the earth's history by a partial pressure of neutrons in the mass, of sufficient concentration and sufficient translational energy to give reaction velocities adequate to maintain equilibrium. Without going into the detail of these reactions at this stage, if the reaction velocities were comparable for the nuclei of different masses, the same freezing-in parameter would be found, *with respect to this reaction* (equation (2) above), for the whole of the periodic system. On the other hand, exceptionally sluggish nuclear reactions of this type would have greater freezing-in parameters, and exceptionally speedy nuclear reactions would have smaller freezing-in parameters (temperature and partial pressures). A complication arising from radioactive nuclei in equilibrium with stable nuclei is discussed below.

§2. STATISTICAL CALCULATIONS OF FREEZING-IN PARAMETERS

Although the equilibria between nuclei, photons, and elementary particles at high temperatures may be governed by statistics which differ substantially from those which hold at ordinary temperatures (Wataghin, 1944; Lattes and Wataghin, 1946), many of the uncertainties cancel out in the particular problem examined in this paper. In reversible processes of isotope formation, such as

$$ {}^{m}_{n}A + {}^{1}_{0}n \rightleftharpoons {}^{m+1}_{n}A + \Delta H_1, \qquad \ldots\ldots(3)$$

$$ {}^{m+1}_{n}A + {}^{1}_{0}n \rightleftharpoons {}^{m+2}_{n}A + \Delta H_2, \qquad \ldots\ldots(4)$$

equilibrium constants may be defined in the usual way, e.g.

$$K = [{}^{m+1}_{n}A]/[{}^{m}_{n}A]\,.\,[n], \qquad \ldots\ldots(5)$$

where the terms in square brackets refer to the thermodynamic activities of the appropriate species. These equilibrium constants will be related to the heat evolved, ΔH_1, etc., according to standard formulae (cf. Ubbelohde, 1937)

$$\log_e K = -\frac{\Delta H_0}{\boldsymbol{k}T} - \frac{1}{\boldsymbol{k}T}\int \Sigma C_p\, dT + \frac{1}{\boldsymbol{k}}\int C_p \frac{dT}{T} + I'. \qquad \ldots\ldots(6)$$

This expression may be simplified, and freed from a number of experimental unknowns, as follows:

In place of $\log_e K$, we may write $-\log_e [n] + \log_e r$, using (5), and taking the ratio of the thermodynamic activities

$$[{}^{m+1}_{\;\;n}A]/[{}^{m}_{n}A] = r$$

under the conditions of freezing-in, as given by the ratio of concentrations r in which isotopic nuclei are found to be stabilized under current terrestial conditions. Owing to the close similarity in each pair of nuclei considered, this approximation appears to be justified, even if the thermodynamic activity itself differs largely from the partial pressure, under the conditions of freezing-in.

In place of ΔH_0, the energy which would be absorbed in the nuclear transformation if it occurred at 0° K., the Einstein rest-mass equation gives

$$\Delta H_0 = c^2\,[{}^{m+1}_{\;\;n}A - {}^{m}_{n}A - m_n], \qquad \ldots\ldots(7)$$

where the terms inside the brackets refer to the atomic masses of the particles. It is not necessary to know the exact mass m_n of the neutron for what follows, but the difference in mass between two isotopes differing by unity in mass number must be known with high accuracy. It may be written ΔA, so that

$$\Delta H_0 = c^2\,[\Delta A - m_n]. \qquad \ldots\ldots(8)$$

Using as mass scale for ΔA, ${}^{16}\mathrm{O} = 16$, the unit of mass will be $1{\cdot}66 \times 10^{-24}$ (Aston, 1942), and inserting for

$$\boldsymbol{k} = 1{\cdot}379 \times 10^{-16}\ \text{c.g.s.},$$

$$c^2 = 9 \times 10^{20}\ \text{c.g.s.},$$

$$\Delta H_0/\boldsymbol{k} = 1{\cdot}08 \times 10^{13}\,[\Delta A - m_n].$$

A quantity which is also required has the value

$$\Delta H_0/2{\cdot}303\boldsymbol{k} = 4{\cdot}7 \times 10^{12}\,[\Delta A - m_n].$$

At the freezing-in temperatures it seems unlikely that the specific-heat terms will involve anything but translational energy, since nuclear excitation is probably incompatible with freezing-in, and the moment of inertia of the nucleus I can be neglected.

[Writing $\Theta_{\text{rot}} = \boldsymbol{h}^2/8\pi^2 I\boldsymbol{k}$, if the radius of the nucleus is of the order of 5×10^{-13} cm., the characteristic temperature Θ_{rot} for the excitation of rotation is of the order $2{\cdot}5 \times 10^{13}/A$° C. where A is the mass number.]

Thus ΣC_p for two particles condensing to one will be equal to $-\frac{3}{2}\boldsymbol{k}$, i.e.

$$-\frac{1}{\boldsymbol{k}T}\int \Sigma C_p\, dT + \frac{1}{\boldsymbol{k}}\int \Sigma C_p\, dT/T = \tfrac{3}{2} - \tfrac{3}{2}\log_e T.$$

This expression will not be strictly applicable if equilibrium was frozen-in at high densities and very high temperatures, but it is quite adequate for testing whether the same freezing-in parameters apply to the different atoms of the periodic system.

By a standard result, the constant $I' = \Sigma i_m$, where the constants i_m correspond to the entropy constants for the various particles

$$i_m = \log_e \frac{(2\pi m k)^{3/2} k g_m}{h^3}.$$

Apart from constants, i_m is a function of the mass and the spin multiplicity g_m of the nucleus. For the simple equilibrium considered,

$$I' = \log g_{m+1}/g_m + \log C$$

where the term $\log C$ depends only on the neutron, and universal constants, apart from acorrection for the fact that the ratio of the atomic masses $\log ({}^{m+1}A/{}^{m}A)$ differs from unity. Except for hydrogen, this correction can be neglected compared with the other uncertainties still present in the expression.

Inserting these simplifications into (6), rearranging, and using common logarithms, the freezing-in parameters may be calculated from the equation

$$\log_{10} r = -4{\cdot}7 \times 10^{12} \Delta A/T + \log_{10} g_{m+1}/g_m + K_e \qquad \ldots\ldots(7)$$

where the term

$$K_e = 2{\cdot}303 \log_{10}[n] + 2{\cdot}303[-1{\cdot}08 \times 10^{13} m_n/T + \tfrac{3}{2} - \tfrac{3}{2} \log_e T] + \log_{10} C$$

is independent of the masses ${}^{m+1}_{n}A$, ${}^{m}_{n}A$, and is constant for all nuclei with the same freezing-in parameters T and $\log_{10}[n]$.

The constancy of these parameters may conveniently be tested graphically, by plotting experimental values of $-\log_{10} r$ against experimental values of ΔA. At present, these are available with sufficient accuracy in the case of a limited number of atoms only (Pollard, 1940; Okuda *et al.*, 1941).

In this plot, reproduced in figure 1, it is convenient to include data for isotopes differing by mass number 2 (triangles), as the intermediate isotope is sometimes rare. Data for isotopes differing by mass number 1 are expressed as circles. The thermodynamic equation for the mass difference 2 is readily obtained by adding the equations for the two steps (3) and (4) above, and making the appropriate adjustments to the constants. Provisionally, until more general information is available about nuclear spins, no allowance has been made in the plot for the ratio g_{m+1}/g_m being different from unity in some cases. If j_m is the spin of the isotope of mass number m, the order of correction to $\log_{10} r$ is $\log_{10} \frac{(2j_{m+1}+1)}{(2j_m+1)}$, which will normally be considerably less than $\pm 1{\cdot}00$, and so does not obscure the main conclusions from the diagram. Examination of this shows that there is an undoubted general correlation between the mass (or energy) change on forming one isotope from another and the abundance ratio as at present stabilized in the earth's substance.

On the basis of equation (7) above, the freezing-in temperature which corresponds with the heavy line drawn in the diagram has a value given by

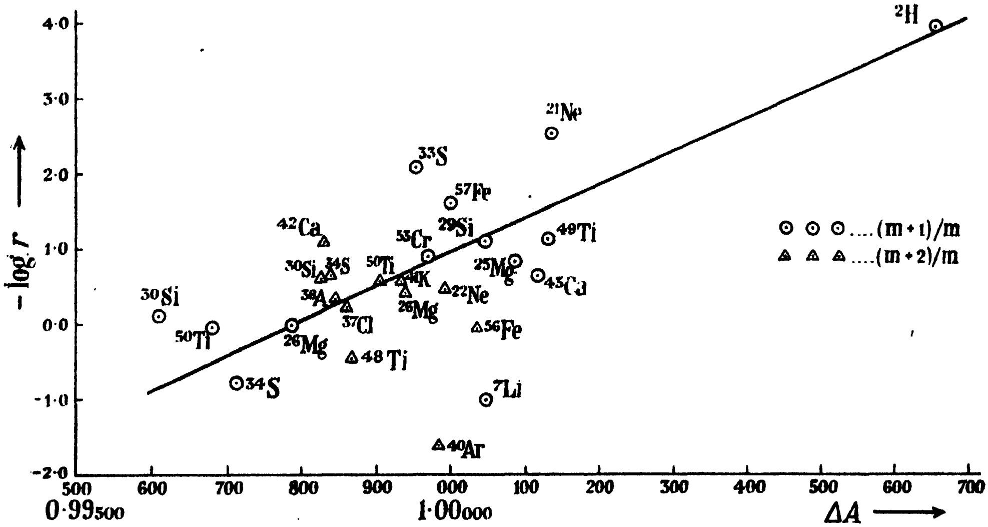
$$\Delta(\log r)/\Delta A \simeq 1{\cdot}0/0{\cdot}0023 = 4{\cdot}7 \times 10^{12}/T$$

or

$$T \text{ (freezing-in)} = 10^{10}\ {}^\circ\text{c}.$$

The fact that this freezing-in temperature appears to hold within fairly close limits for a considerable number of atomic species is perhaps rather surprising. It suggests that the reason for freezing-in may have been a falling off in the term log [*n*] for the partial pressure of neutrons, which would affect many of the reactions in the same way, rather than inadequate activation energy of the neutrons, which would be expected to occur rather sooner for the larger values of ΔA.

Particular interest is attached to the exceptional values of the freezing-in parameters. At the present stage of our information about nuclear spins it is probably advisable to treat values of $\log r$ within ± 1·00 of the freezing-in curve as not yet clearly differentiated. If lines parallel to the curve are drawn above and below it, at this distance, this leaves certain isotopes of Si, Ca, S, and Ne with exceptionally high values of $\log r$, and of Ti, Fe, Ca and Li, Ar with exceptionally low values of $\log r$. The interpretation of these divergences is of considerable importance. This paper will deal with one or two points only:



(i) A trivial explanation for some of the divergences is that experimental values of *r*, or more probably of ΔA, are incorrect. Further work on accurate measurements of mass differences should clear this up in due course.

(ii) Under the conditions of freezing-in, an appreciable proportion of the nuclear species in the earth's mass may have been radioactive, i.e. capable of undergoing spontaneous disruption even in the absence of bombardment by other particles. For such species, the values of *r* would continue to change even after freezing-in of the equilibrium processes. This possibility should not be beyond the reach of eventual experimental verification, since the energy difference between those stable nuclei which are found in abnormally large amounts, and the parent radioactive nuclei, may be obtained from atomic reactions studied in the laboratory. The appropriate value of ΔA could then be inserted in the plot, or conversely, if the parent nucleus is known, the value of ΔA could be approximately calculated from the plot, a procedure which may be of value in some cases.

(iii) In the absence of other explanations, outstanding exceptions such as the ratio $^7\mathrm{Li}/^6\mathrm{Li}$ would point to nuclear reactions capable of maintaining this equilibrium at lower values of neutron activity than the bulk of the processes responsible for the maintaining of detailed balancing in nuclear equilibrium. It is of considerable interest to examine how far such reactions can be verified under laboratory conditions.

(iv) It will be noted that the form in which equation (7) has been stated deliberately avoids a number of difficulties, whose discussion is of great interest, but which introduce unnecessary uncertainties in the conclusions. For example, the thermodynamic activity of the neutrons under freezing-in conditions can be calculated from the value of the intercept of the curve in figure 1, but this involves further assumptions, which will not be dealt with here. It is interesting to observe, nevertheless, that an increase in the thermodynamic activity of the neutrons under freezing-in conditions would shift the whole curve upwards (by increasing the intercept of the $-\log r$ axis, algebraically), whereas a change in the freezing-in temperature would merely affect the slope of the curve.

(v) In calculating the numerical value of the freezing-in temperature, no allowance has been made for possible changes in the scale of atomic energies with time. If differences in energy between isotopes were substantially smaller at the time of freezing-in than they are calculated to be from currently observed mass differences, this would lower the estimated freezing-in temperature, but it would not affect the general use of the plot as a test for abnormal nuclear transformations. [Cf. references to the work of E. A. Milne and colleagues in Johnson, (1945).]

REFERENCES

Aston, 1942. *Mass Spectra and Isotopes* (London : Arnold and Co.).
Chandrasekhar and Henrich, 1942. *Astrophys. J.*, **95**, 288.
Harkins, 1917. *J. Amer. Chem. Soc.*, **39**, 856.
Johnson, 1945. *Time, Knowledge and the Nebulae* (Faber and Faber).
Lattes and Wataghin, 1946. *Phys. Rev.*, **69**, 237.
Pollard, 1940. *Phys. Rev.*, **57** 1186.
Okuda *et al.*, 1941. *Phys. Rev.*, **59**, 104; *ibid.*, **60**, 690.
Seaborg, 1944. *Rev. Mod. Phys.*, **16**, 1.
Wataghin, 1944. *Phys. Rev.*, **66**, 149.
v. Weizacker, 1938. *Phys. Z.*, **38**, 632.

RUTHERFORD AND THE MODERN WORLD

By M. L. OLIPHANT, F.R.S.

The Third Rutherford Memorial Lecture, delivered 7 October 1946

§1. RUTHERFORD THE MAN

Can we attempt to assess the impact of Rutherford and his work upon the science and life of the present time?

Perhaps it is too early to form any balanced judgment of the value of his work and of his influence on others, but it should be possible, even for one who cannot be dispassionate about a man whom he revered as a scientist and loved as a man, to make some timely remarks upon Rutherford's major contributions to the scientific life and outlook of to-day.

Before considering any details it is instructive to measure his greatness in figures concerning one aspect of the release of atomic energy—an achievement of mankind based almost entirely on the work of Rutherford, his colleagues and his students. During 1945, the electricity undertakings of this country generated 37,281 millions of kilowatt-hours of electric power. If the steam plant employed had the high overall efficiency of 30 per cent this would require the burning of about 18 million tons of coal. If the inherent possibilities of atomic energy can be realized in practice, and many of us are convinced that they can, the amount of "atomic fuel" required, if derived from uranium or thorium, would not be much more than 5 tons. In order that the cost of electricity should be the same for both types of fuel, the prime cost of "burning" nuclear fuels could be 3 million times as great as that of burning coal. It is not unreasonable to believe that without Rutherford such great possibilities would not have been realized for a very long time.

Rutherford's faith in nuclear physics as a field of human endeavour was not shared by all. He was often criticized for failing to work, and to train his students, in the useful aspects of physics which had already found practical application. His greatness is apparent in his steady pursuit of the frontiers of physical knowledge, leaving to lesser men the more obvious work of consolidating knowledge, the broad outlines of which had already been explored. He recognized good work in any branch of physics or of any of the other divisions of science, but he had a special love for his own chosen subject, which he called "A Tom Tiddler's Ground", where anything might turn up and where preconceived ideas and theories often toppled to the ground as new experimental facts were discovered.

§2. RUTHERFORD'S DISCIPLES AND THE WAR

Real fortune favours few. This is as true in physical science as in other walks of life. Those who were able to work as pupils and colleagues of Lord Rutherford are to be numbered among those on whom fortune has smiled. Many who worked with him and knew him well have tried to visualize his reaction to the terrible possibilities which the new things in science make probable, to the intrusion of secrecy into pure science and to the growing demand for a special place in society, with proportionally greater income, for scientists of all kinds. Although we accept the fruits of the new spirit and of the new regard of the world for the scientific wizards whom it fears, at least we know that his view was right and that ours is wrong. For him our compromise would have been impossible.

During the last hundred years there has been a continual succession of great men in British physical science, from Faraday and the giants of the Victorian era to Rutherford himself. Today there is no great figure, no one man who by his work, his teaching and his example, can exert the same influence on science as a whole or upon the government of the country. Perhaps the day of the great individual in physics is past and teams of lesser men will progress faster and faster. However, the lessons of the past show that the landmarks of discovery were the product of rare genius and it is difficult to believe that this will not be so in the future.

The development of atomic energy has made use of the accumulated capital represented by the work of Rutherford. Great as is the material achievement, it has been in fact development work—what is called applied physics—and no new discovery has been made in fundamental physics which is of the first or even of the second magnitude. The last six years have been years of stagnation in pure physics, and there is much leeway to make up if there is to be a worthwhile physics in the future. The present generation of physicists has had no contact with the atmosphere of a research laboratory. It knows only the feverish atmosphere of development for an urgent application for war purposes or of a study of phenomena for a material end. Its salaries have often been as high as Rutherford ever earned, and its idea of personal progress has been to secure elevation to a position commanding authority and increased pay. Its ability to recover from this diseased state and to recapture the spirit which animated Rutherford will be a measure of the ability of this country to play a great part in the physical science of the future.

When the war began there were two great projects where physicists could play a major part. The possibilities of obtaining atomic energy through the fission process had been made clear by the discovery, by Joliot and his co-workers, that neutrons were emitted. The revolutionary idea of an atomic bomb of unprecedented power was envisaged early in the war by Frisch and Chadwick. However, the problem of separating the isotopes of uranium, which seemed essential, was obviously one demanding long-term development and involving much uncertainty. The imminent dangers threatening this country made the alternative problem of radar defence more attractive to many physicists. In both these fields Rutherford's pupils brought his methods of direct attack to bear, and made great strides very rapidly because they were able, at all times, to build on fundamental principles and were unhampered by standards of practice.

Chadwick, to whom Rutherford's leadership in nuclear physics had naturally descended, played a part in the development of atomic energy which will be fully appreciated only when the whole story is told. In his handling of men who worked with him, in his delicate task of relationship with the U.S.A., his legacy from Rutherford was clearly apparent. Perhaps he alone of the British scientists who served in the war preserved throughout that clarity of physical insight and feeling for fundamentals which was characteristic of Rutherford. Most of us found ourselves greatly affected by the scientific compromise which seemed essential to progress and as a result emerged from the war more interested in gadgetry and quick results than in the intellectual side of science.

The migration of Rutherford's students into radar was a surprising feature of the early years of the war, and here they played a prominent part. Men like Cockcroft, Lewis, Dee, Skinner, Ratcliffe, and many others working with Watson-Watt and Rowe, together saved Great Britain from defeat and contributed greatly to her victory. They introduced into the work a totally different atmosphere, for they worked with their fellows as equals and believed in detailed discussion, even with the junior workers, at all stages. The story of what they did and how they achieved it has been told again and again, but one and all they acknowledge that they merely handed on or used the methods which they had learnt from Rutherford and Chadwick in the Cavendish Laboratory.

Blackett, after playing a part in radar, moved on to the Admiralty, where he brought to a high pitch of perfection the new art of operational research—the application to actual military operations of the methods of scientific reasoning.

With the help of Fowler, E. J. Williams and Bullard, he was largely responsible for the defeat of the U-boats at a crucial period in the war.

In these fields and in many others Rutherford's pupils assured the predominance of this country in science applied to war. Tizard, a close personal friend and great admirer of Rutherford, led a mission to America which handed over to its scientists the fruits of our experience and invention, so that when they joined us in the war they also shared the new applications of scientific method and practised these for themselves.

It is disappointing that the full fruits of this influx of Rutherford's spirit into the government and service establishments are not being retained. It was unlikely that many of the best physicists would choose to remain in employment where secrecy and restrictions applied to their work, but such splendid establishments as T.R.E. might have been preserved as national laboratories, possessing a great deal of freedom from day-to-day routine, with great profit to our country.

§ 3. RUTHERFORD AND THE EMPIRE

Rutherford never forgot that he was a New Zealander or that his early work was done in Canada. He was always sympathetic towards students from the Dominions who came to work with him. His laboratories in Manchester and Cambridge became the Mecca of those overseas students who were lucky enough to go abroad to study, and particularly for those physicists who came to England under the auspices of the Exhibition of 1851. Here they found Mr. Evelyn Shaw, the Secretary of the Commission, ever ready to help them carry on, and in Rutherford they found the kindly sympathy in personal matters, combined with hard drive and an impassioned belief in experimental science, which brought out all that was good in them. Under his guidance, and that of Chadwick, some of us who must have been unpromising material became reasonably competent physicists who, whatever their limitations in fundamental intuition, at least try to keep alive the spirit of his approach. Some have remained in this country, others have returned to the Dominions, and have enriched scientific life there out of all proportion to their numbers. Marsden in New Zealand and Schonland in South Africa are responsible for their countries' science as a whole, while in the Universities and other scientific institutions Rutherford's students hold a surprising proportion of important posts. In the Dominions, as in this country, these men played outstanding parts in their countries' efforts when war came.

Rutherford's Dominion birth and sympathies gave him also a passionate belief in the British Commonwealth of Nations. He seemed always to think of the Commonwealth as a unit, the existence of which, and the predominance of which, he never questioned. This reacted not only on men from overseas but also on his English students and colleagues, so that among them there is a greater appreciation of the problems of the Commonwealth than is common among the citizens of any part of the Empire. Yet he was a true internationalist, ready to give credit for achievement whatever the origin, colour or creed of the worker. At the time of the German persecution of the Jews he was foremost in his defence of their claim on us and worked indefatigably for their cause. He gave to many a home in his laboratory and helped others to secure positions abroad.

Rutherford's faith in the British Commonwealth and in international co-operation became the faith of many of his students. It is difficult, for instance, for these men to understand why in great projects, such as atomic energy, the problem is not tackled co-operatively within the Commonwealth, whose different parts could contribute in large measure to the success of a venture which may mean far more to them in its peaceful aspects than it will ever mean for the United States.

Rutherford admired greatly the experimental genius of Faraday and was intimately familiar with his diary. In his opinion Faraday was the greatest of experimental physicists. When a full assessment is made of Rutherford's own work and of his influence on physics I think it likely that he will rank with Faraday as an experimentalist, while the inspiration he gave to his students and collaborators will place him above Faraday in the sum total of his contributions to science. Faraday's work laid the foundations of electrical engineering; Rutherford's is the corner-stone upon which is based the exploitation of atomic energy.

§4. NUCLEAR PHYSICS

Rutherford was not afraid to put forward a hypothesis which was helpful in explaining experimental results because further evidence might prove it wrong. He did not hesitate to publish experimental data which, because of the nature of the problem, might turn out to be incomplete or wrongly interpreted. Yet his work, and that of collaborators in his laboratory, was singularly free from those hasty and misleading conclusions which have sometimes been published in the insane struggle for priority. His earlier work was extraordinarily painstaking and complete. His discovery, and subsequent investigation, with Chadwick, of artificial transformation by α-particle bombardment has been amplified, but in no case disproved. When, in his laboratory, Chadwick discovered the neutron and Cockcroft and Walton first observed transmutation by artificially accelerated particles, the evidence given was complete and satisfying and started a whole train of fresh work throughout the world. This ability of Rutherford to transmit to others his care and thoroughness, as well as his enthusiasm, has been responsible for the upsurge of great work in nuclear physics in this country and abroad.

Rutherford's strong personal association with Niels Bohr, from the Manchester period onwards, contributed to his proper but balanced appreciation of the place of theoretical physics in the advance of the subject. He refused to be bluffed by the occasional enthusiast who felt that the solution of the appropriate wave-equation could give the answer to all problems of physics. To the end of his life he had an almost fanatical belief in the power of the experimental method. The discovery of such unsuspected phenomena as the fission of uranium and thorium, leading to the atomic bomb, was a striking example of the correctness of his point of view. In fact theoretical advance is due as much to the experimental discovery of new facts and new laws as to advances in theoretical technique, and at present fundamental theory is waiting upon experimental results which would have been available but for the war. Little progress is to be expected in the theoretical physics of the nucleus until data are available on the scattering laws for protons and neutrons at energies in excess of 100 Mev.

Rutherford believed that because of the inevitable lag between academic discovery and practical application it should be possible to foresee, to a useful extent, the general trend of application of science to industry, and in that way to provide in advance for new things, thus avoiding dislocation of economy and reactionary fighting against progress in technology. In one of his rare speeches in the House of Lords he advocated the setting up by the government of a "Prevision Committee", consisting of scientists and others, who could advise on economic changes which might result from existing scientific knowledge. So far as I know this very practical suggestion has not been implemented, but clearly it will become more necessary as science is applied more vigorously to our economy.

An academic atmosphere, with complete freedom to follow whatever paths of investigation seem desirable or necessary, was, according to Rutherford, an essential for progress in fundamental science. He felt it to be a grave mistake for a man with high abilities in academic physics to be tempted into industrial or State laboratories. I remember an occasion when the director of scientific research of one of the service departments asked Rutherford whether he had one or two good men who could join his division. Rutherford's immediate reply was "Look here, So-and-So, I know the conditions in your place; do you seriously think if I had a good man I would send him to you?" Perhaps such an attitude does not help to improve the standards of the scientific services, but it is certain that the men he refused to send at that time had a far freer hand and made a greater contribution to victory when they joined the Services with fresh minds after war had begun than would have been possible under the conditions applying to Service establishments at an earlier date. When I became an Assistant Director of Research in the Cavendish Laboratory, Rutherford talked to me about the choice of investigations for the research students. "You know, Oliphant, in this game it is rather important to choose the right experiments to do, but it is even more necessary to know when to stop." He added that in industrial and government laboratories men were often assigned to a problem, or class of problem, and had to work in that field till they retired. He believed that no man could make creative contributions to a subject through a particular line of attack for more than a few years. After that he became stale. Yet, despite his reluctance to see the really original mind in physics leave the fundamental for the applied side of the subject, he was no intellectual snob.

He recognized and admired achievement in applied physics and gave credit to industry for the tools and devices its work provided for his own experiments.

I have indicated already that the development of atomic energy is likely to rank in the modern world with the birth of electrical engineering in the 19th century. Recent very conservative reports by engineers in America give a very optimistic picture of the possibilities of providing industrial power on an economic basis by "burning" uranium in a modified form of the relatively inefficient type of "pile" operating at Hanford. They conclude that on the same assumptions with regard to amortization of capital charges and allowance for running-cost, electricity generated from atomic energy would cost in the U.S.A. about 30 per cent more than electric energy derived from coal, and they point out that coal costs are rising steadily whereas the cost of atomic energy will undoubtedly decrease as development proceeds. On this basis electrical energy could already be derived from uranium in this country at a cost less than that from coal-burning stations; some of us believe that with reasonable development the cost will be considerably less.

Devons has pointed out that it is probably unimaginative to think of atomic energy merely as a substitute for other forms of energy. The great difference in kind and in concentration of the energy will mean that it will be applied in totally new ways for totally new purposes. The atomic age will arise not so much from a replacement of existing fuels by fissile materials as from new processes which now become possible.

It helps to understand the greatness of Rutherford if we examine the outstanding events, the landmarks, in the unravelling of our understanding of the nucleus, which has culminated in the release of atomic energy. In that way we can see clearly how firmly it all rests on his work; on the intuition of his genius.

The nature of radioactive change

After the discovery of radioactivity by Becquerel and the isolation of radium by the Curies, the further development of our knowledge of radioactive change came about almost entirely from Rutherford's work. The natures of the three types of radiation were recognized by him and named α, β, and γ radiations. Brilliant experiments proved that the α particles were charged atoms of helium. The nature of the ionizing effects of the radiations was developed and the energies deduced. With Soddy he put forward the conceptions of radioactive series and of isotopes.

The nuclear model of the atom

The α particles were always special favourites of Rutherford. He was intensely interested in all their properties and spoke of them as though he knew their colour and even their characters and idiosyncracies. It was natural that he should investigate in detail what happened when these particles passed through matter. For this purpose he developed the use of the scintillation method for detecting single α particles, a method discovered by Crookes, and with Geiger he invented the electrical counter which has played such a large part in modern work. The deviations from straight paths which were occasionally observed led to the careful investigations of scattering and of the scattering laws. On the assumption that the forces between the charged α particles and other atoms were electrical in nature, Rutherford and his colleagues were able to show that the closest distance of approach of the energetic α particles to the centre of electric force in the nuclei of the struck atoms was far smaller than the radius of the atom itself. This led to the modern picture of the atom as a central core, or nucleus, in which is concentrated practically the whole of the mass, carrying a positive charge equal to the atomic number of the atomic species, and surrounded by the number of orbital electrons necessary to make the atom neutral.

Electrons, rotating about the nucleus, were not stable, on the hypothesis that the laws of mechanics governed their motions, as they would radiate energy and fall into the nucleus. Bohr, at that time working with Rutherford in Manchester, found a solution to this problem by the bold conception of stationary quantum states, and the nuclear model of the atom was securely launched.

Artificial transmutation of atoms

During the work on the scattering of α particles it was observed that fast protons were produced. These protons could have arisen from collisions between the

α particles and hydrogen present as an impurity in the bombarded substances. The transfer of momentum which occurs in such collisions would give to the protons a definite maximum energy, actually 16/25 of the energy of the α particles. Examination of the ranges, and therefore of the energies of the ejected protons, showed that in most cases they did arise in this way. However, in the case of nitrogen Rutherford observed that protons were produced with energy considerably greater than this maximum. He concluded that the only possible explanation was to assume that the nitrogen had undergone a nuclear transformation in the collision, that the α particle remained inside the struck nucleus and a proton was ejected, the nitrogen being transformed into oxygen:

$$^{14}_{7}\mathrm{N}+^{4}_{2}\mathrm{He}\rightarrow{}^{17}_{8}\mathrm{O}+^{1}_{1}\mathrm{H}.$$

In this reaction it was assumed that the proton carried away the excess energy available in the reshuffling of the nuclei.

This was the first observation of the artificial transformation of one element into another. Blackett, in Rutherford's laboratory, confirmed these assumptions by observing the reaction in the expansion chamber, showing that momentum was conserved but that energy was not.

An extensive series of observations was then made by Rutherford, with Chadwick and others, and it was shown that several of the light elements, the nuclei of which could be penetrated by the α particles available, could undergo similar transformations. Experiments were afterwards carried out elsewhere, notably in Vienna, but in general the results were not reliable, and the evidence accumulated in the Cavendish Laboratory remained unchallenged.

In the course of this work it was observed that beryllium when bombarded with α particles gave rise to a very penetrating type of radiation which was thought to be high-energy gamma-radiation. This radiation was investigated in some detail in the Cavendish Laboratory, in Germany and in France, and it was found to be difficult to reconcile its properties with those of other forms of γ radiation. Joliot and his wife, the daughter of the discoverer of radium, found that the radiation from beryllium was able to project protons from hydrogen-bearing substances in a manner similar to the projection of electrons by x rays in the Compton effect, but they found it very difficult to devise a picture of the process which would obey the laws of conservation of momentum and energy without assuming a prohibitively large energy for the γ ray. It was difficult also to account for the large probability of ejection on the assumption that the radiation was electromagnetic in character.

Discovery of the neutron

Chadwick returned to this problem in the Cavendish Laboratory, and in 1932 discovered that the radiation from beryllium was able to project atoms of all kinds, the maximum kinetic energy of which varied in a regular way with the mass of the struck atoms. He concluded that it was impossible to account for these results on the assumption that the radiation was electromagnetic in character, and he put forward the hypothesis that beryllium, when bombarded with α particles, emitted a new kind of particle which carried no electric charge. He was able to calculate the mass of these " neutrons " from the variation of the energy of various atoms which recoiled in collisions, and showed that it was about the same as the mass of the proton

In his Bakerian lecture to the Royal Society in 1920, Rutherford discussed the properties of a neutral particle of mass unity. "Under some conditions it might be possible for an electron to combine.... closely with the H nucleus, forming a kind of neutral doublet. Such an atom would have very novel properties. Its external field would be practically zero, except very close to the nucleus, and in consequence it would be able to move freely through matter. Its presence would probably be difficult to detect by the spectroscope, and it may be impossible to contain it in a sealed vessel. On the other hand it should enter readily the structure of atoms, and may either unite with the nucleus or be disintegrated by its intense field, resulting possibly in the escape of a charged H atom or an electron, or both... The existence of such atoms seems almost necessary to explain the building up of he nuclei of heavy elements...."

It is clear that Rutherford understood to a remarkable degree the importance of a neutron as a fundamental particle in the structure of nuclei, and he knew what properties such a particle must possess. It was natural that an intensive search should be made in the Cavendish Laboratory for evidence supporting this hypothesis, but this search was unsuccessful until the advances in counting technique by electronic methods which were due to Wynn-Williams, also working in the Cavendish Laboratory, made it practicable to detect the nuclei recoiling elastically from neutron collisions.

With this background of expectation, Chadwick's discovery of the neutron followed naturally from the years of intense study of nuclear phenomena which Rutherford and he had spent. Chadwick's paper describing the discovery is a model of completeness. There remained no doubt whatever about the existence of the new particle or of its properties.

Transmutation by artificially accelerated particles

Rutherford became interested in the possibilities of producing fast particles with which to bombard nuclei, for he realized that the energy, kind and number of bombarding particles available from radioactive substances were very limited. New information might be obtained by using particles other than α particles, and effects which were too weak to observe might become appreciable if much stronger sources of bombardment were available. Accordingly Allibone, and later Cockcroft and Walton, developed in the Cavendish Laboratory techniques for producing high voltages and for applying them to evacuated accelerating tubes. In particular, Cockcroft and Walton accelerated protons derived from an electric discharge through hydrogen and bombarded targets of the light elements. Although the potentials which they had reached were only 400000–500000 volts, it was felt worth while to try since, according to the newly developed wave-picture of the penetration of particles into nucleus, detectable effects might be anticipated for the lightest elements.

These experiments gave a successful result in 1932. It was found that the penetration of protons into lithium gave rise to a copious emission of α particles. This important observation has had a tremendous effect on physics, for it started the whole modern era of nuclear physics. The reaction may be written:

$${}^{7}_{3}\mathrm{Li} + {}^{1}_{1}\mathrm{H} + \text{k. e. of proton} \longrightarrow {}^{4}_{2}\mathrm{He} + {}^{4}_{2}\mathrm{He} + \text{energy},$$

and if we substitute the known masses of the atoms involved, including the mass-equivalents of the energies, we find:

$$7{\cdot}0182+1{\cdot}0081 \rightarrow 4{\cdot}0039+4{\cdot}0039+0{\cdot}0185.$$
$$8{\cdot}0263 \rightarrow 8{\cdot}0263.$$

This remarkable equality proved quantitatively for the first time the exact validity of the Einstein relation $W=M.c^2$. Several other light elements were shown to undergo transmutations when bombarded by protons. Rutherford and others in the laboratory followed up this pioneering work by a careful analysis of the effects produced by bombarding substances, especially deuterium itself, with ions of deuterium, or deuterons. This last series of observations was carried out with less than one-fifth of a cubic centimetre of heavy water provided from his first production by Professor G. N. Lewis of California.

This work was rapidly extended in other parts of the world, especially by E. O. Lawrence in California, using the powerful cyclotron method of acceleration which he had developed. It became the fashion in physics to bombard atoms, and all over the world equipment grew up to enable experiments of this sort to be carried out. However, although many more reactions were studied in detail and far more powerful effects were observed elsewhere, the pioneering work in Rutherford's laboratory remained the supreme example. No discoveries of importance approaching these early observations were made elsewhere.

Artificial radioactivity

The Joliot-Curies, in Paris, had discovered that certain substances could be rendered radioactive by bombardment with α particles. The radioactive materials were unstable isotopes of the normal elements, which transformed by emission of positive or negative electrons into stable isotopes of neighbouring elements. Cockcroft and Walton showed that these radioactive forms of the elements could be prepared by their method of bombardment with protons or deuterons, and this rapidly became the standard method for preparing strong sources, especially where cyclotrons were available. These artificially radioactive substances can now be produced in almost unlimited quantities as a by-product of atomic energy, and will soon be available in this country, as in America, for use in chemistry, biology and medicine as "indicators" of the movements and history of any element in any compound.

Transformations produced by neutrons

Feather, in the Cavendish Laboratory, first showed that neutrons could produce nuclear transformations in exactly the same way as protons or α particles, but the major early work on transformations produced by neutrons was done by Fermi and his collaborators in Rome. In particular it was shown that neutrons could be slowed down to thermal energies by multiple collisions with hydrogen in such substances as water or paraffin wax, and that these slow neutrons were particularly effective. They were able to enter most nuclei and were often captured there, producing unstable radioactive isotopes of the same element. In some cases the probability of capture was extremely high due to the existence in certain nuclei of energy levels enabling resonance transitions into the system by neutrons with small kinetic energy.

Fermi examined the reactions produced in most elements. The neutron, being uncharged, is able to enter heavy nuclei of high atomic number as easily as those of small mass and charge, so that in contrast with accelerated protons and deuterons strong effects were observed with the heaviest elements. His observations with uranium led him to believe that by electron emission after neutron capture, elements of greater nuclear charge than uranium, the so-called transuranic elements, could be formed, but some of the evidence was not very satisfactory.

These observations were followed up in Berlin by Hahn, who had been one of Rutherford's earliest research students in Montreal. Using chemical methods of separating the radioactive products of the reaction of slow neutrons with uranium, he was able to show, in 1938, that these were often elements of medium atomic weight, like barium and strontium. He therefore put forward the hypothesis that uranium, after absorption of a neutron, could undergo a new type of transformation by splitting into two approximately equal parts—a "fission" process, which was accompanied by an enormous release of energy. This hypothesis was immediately confirmed in Copenhagen and in America. Shortly afterwards it was shown by Joliot and his collaborators that in the fission process several neutrons were emitted. At once the possibility of producing a chain process became apparent, and the search began all over the world for methods of achieving such a divergent process, in order to produce both industrial power and atomic bombs.

Surely, after consideration of the process by which our knowledge of nuclear physics has advanced to the stage where atomic energy is available, we must admit that of all men Rutherford must stand out as the pioneer of the new age. His was the genius which gave it birth.

The future of nuclear physics

Nuclear physics, as a subject for academic research, faces a serious crisis in all countries. The grave military implications of atomic energy are bound to mean restriction on the field of activities in nuclear physics which can be permitted without control and supervision. It is fortunate that the sections of the subject which need special supervision are less those dealing with the frontiers of our knowledge than those concerning the detailed accumulation of information about particular types of nuclear reactions. This means that academic nuclear physics can still be carried on with some considerable freedom in the more fascinating fields of the subject provided only that the law relating to control of atomic energy is wisely interpreted and administered.

§ 5. CONCLUSIONS

Rutherford's greatest contributions to our knowledge of the structure of matter came from his intensive investigations of the scattering of α particles—his use of these atomic projectiles as probes with which to examine the fields of force within the atom and even within the nucleus itself. The power and elegance of this method is obvious, and it is clear, too, that much fresh knowledge is to be expected from the use of particles such as protons and neutrons, in place of α particles, especially if much higher energies can be employed. The meson theory of binding for the constituent particles of the nucleus, while providing the best picture we have of the nuclear field, is far from satisfactory, and much more information is required about free mesons, as well as about the laws of force between elementary particles at very close distances of approach. To create a pair of mesons it appears that elementary particles with energies in excess of about 300 Mev. are required, while the scattering of particles of these energies could give information about the laws of force. It seems likely, therefore, that efforts will be made to produce protons and electrons with energies of 1000 Mev. or more. Several methods for achieving this have been suggested, and if the technological difficulties can be overcome we can face with confidence a new era of increasing knowledge of that fertile "Tom Tiddler's Ground" where Rutherford dug so well.

DISCUSSION

on paper by E. W. H. SELWYN and J. L. TEARLE entitled "The performance of aircraft camera lenses" (*Proc. Phys. Soc.*, **58**, 493 (1946)).

Mr. J. W. PERRY. Mr. Selwyn and Dr. Tearle have contributed considerably to photographic optics by this paper. The method employed, considered in relation to the title of the paper, calls for some comment, however, as it is possible to misunderstand its significance if not viewed in the proper perspective. What is presented is not a solution of the problem of photographic lens testing in any fundamental sense, but an empirical bridge to short-circuit important branches of the subject which would have been obstacles to an early fruition of the work. The investigation is, in fact, a typical example of an empirical treatment of an involved subject rendered justifiable by an emergency. Thus one should not be misled by an undoubted success of the work into regarding the results as ultimately valid, for investigation would certainly be needed to put the practical generalizations thus won, on to a fundamental basis and to free them from restrictive conditions and enforced assumptions. For the data upon which the work is based refer only to a certain definite epoch in the course of the development of the photographic objective. They assume certain commercial types of a mixed kind conditioned by different economic and other arguments affecting their form and complexity. They exclude the aid to correction which non-spherical surfaces can contribute. Even the glass situation will have affected one or more of the objectives assumed. Thus there was no fortunate conspiracy of circumstances in the provision of the data upon which Mr. Selwyn and Dr. Tearle had to work.

It is clear from the results that, apart from any possible defects in the photographic plate, the performance of the lens and plate combination suffers from imperfections in the lens, and that with successive improvements in lens design and in methods of manufacture and of test, steady removal of these and approach to the ideal should result in a general improvement in photographic performance, so far as the lens is concerned. But what is the ideal ? That cannot, of course, be learned from an empirical investigation in which varying physical conditions, instead of being isolated for individual study are confused by other unrelated variables coexisting and simultaneously operative. For this the lens must be considered from the point of view of its immediate function and physical effect upon waves of light. The lens testing interferometer provides precisely this information for lenses of any kind, aircraft camera lenses of course included.

The investigation goes considerably beyond this, of course, in order to produce immediately utilizable information and gives results of performance of aircraft *cameras* as a whole. By so doing it adds considerably to our knowledge. If it also thereby loses claim to scientific validity in a fundamental sense, that is by no means a valid criticism, for whereas the combination of physics and empiricism which enters into such an investigation can produce valid results to enable a war to be prosecuted, it obviously cannot take the place of physical studies aiming at an ultimate analysis of the whole process. The results of such a valuable investigation will be all the more valuable when they are actually supported by physical information on the lens performance and on the performance of the photographic plate from the point of view of optical processes.

Mr. G. S. Speak. With regard to the improvements in lenses mentioned by the author it may be of interest to note three outstanding increases in performance which have occurred within the last two years. A 25″ Ross lens is some 50–60% better than would be indicated by the formula, a 50″ Ross telephoto is 40% better, and a redesign of a 36″ telephoto by Wray (original design by Booth) is also 40% better.

Our practice at the Royal Aircraft Establishment when testing lenses photographically is not to give the exposure at each point in the field which leads to maximum resolution, but to expose on axis for this condition and give the same exposure time for points off axis, since this is what happens when the lenses are used in the normal way. For similar reasons the average resolution is calculated over a rectangular picture area instead of a circular area.

Work we have carried out with various test-objects indicate that the probable error of results obtained with a Cobb object is of the order of 7% and that there is no advantage in decreasing the size difference from one group to the next to less than about 5%. Howlett* in Canada obtained similar figures for the accuracy of resolution measurements.

The type of test-object to be used has been discussed many times before. Much obviously depends on the information required from the test, but if the test-object is to bear some resemblance to the details which the lens will be required to photograph eventually, then for testing lenses to be used in air photography an annular test-object proposed by Howlett has numerous advantages, and may be of use in other work.

Finally, I should like to state my opinion that from the point of view of the optical designer resolution tests alone are probably of little use. They should be supported by measurements of aberrations, so that the designer may obtain an idea of the corrections which are most conducive to a good performance. For this reason our reports on lenses for some three years at the R.A.E. have contained full details of resolution tests and measurements of most of the aberrations in the system under test with the idea of placing optical design on a more quantitative basis than it has been in the past.

Mr. G. C. Brock. This work on photographic resolving-power has been described by a previous speaker as " ephemeral ". I think that is probably a sound judgment on a long view and having regard to fundamental aspects, but it does not alter the fact that we have witnessed a very notable advance in photographic optics which would not have occurred without the knowledge accumulated in the course of these tests. In 1940 the Air Photography Research Committee were greatly concerned to improve the resolving-power of aircraft cameras, and realizing that more fundamental investigations would take far too long, sponsored this programme of resolution-tests in the Kodak Research Laboratories. The essential first step in any scientific investigation is to establish some basic facts, and at the time no facts were available with which to answer two questions :—

(*a*) Were lenses or emulsions most in need of improvement ?

(*b*) What was the relation between angular resolving-power and focal length ?

I feel quite strongly that we should not forget how limited was our knowledge of these things in 1940. The results of these investigations by Dr. Tearle and others substituted definite facts for a mass of conflicting opinions and paved the way for the advances made by our opticians in the construction of narrow angle reconnaissance lenses. We should now like to see a similar improvement in wide-angle lenses.

This work has of course made a great contribution to the advances in the general theory of photographic resolving-power, one of the most interesting aspects being the great difference between the photographic and visual resolving-powers of a given lens, and the apparent lack of correlation between them.

---

* Howlett, L. E., *Journal of N.R.C. of Canada* (July 1946).

I should also like to emphasize that in work on air photography, where resolution of ground detail is affected by other things, such as image movement, a knowledge of the resolving-power of the lens/film combination has been an indispensable condition of progress.

The resolution test can be criticized from many angles, and is funadmentally a temporary device which we must hope to see replaced or at any rate extended by some more absolute method involving a measurement of the intensity distribution in the image plane. Nevertheless it has done excellent service over the past few years and we will be returning to it quite often for some time to come.

AUTHORS' reply. We have shown in the paper that the effect on the photographic resolution of the decreasing illumination from centre to edge of the field is very small in the systems investigated, and the method of exposure is therefore of little importance. This result was, of course, unknown at the outset of the work, but was available to the Royal Aircraft Establishment when photographic lens testing was commenced there on a large scale. In lenses where the decrease in illumination with increasing separation from the axis is marked, some modification in the method employed at the Royal Aircraft Establishment would be necessary ; it would, for example, be preferable to give the optimum exposure at some point between the centre and the edge of the field, allowing the axial region to be over-exposed and the peripheral region to be under-exposed.

REVIEWS OF BOOKS

Photoelectric Cells, by A. SOMMER. Pp. 104. (London: Methuen and Co. Ltd., 1946.) 5*s*.

This latest addition to Methuen's Monographs on Physical Subjects gives exactly the kind of information needed for choosing intelligently what kind of emission cell to use as a tool in a particular research, or as a component in apparatus designed for a specific purpose. It also fills the gap between the more advanced text book on photoelectricity, and the standard works on photocell applications which give rather scanty attention to the cell itself. It may therefore serve both as a *vade mecum* for the experimenter in other fields and as a simple introduction to the subject for the non-specialist. It should be noted that it does not deal with barrier layer, or rectifier, photocells.

Many will wish it had been a little larger, and may find in its brevity the reason for a slight distortion of perspective. While the choice of the silver–oxygen–caesium, antimony–caesium and bismuth–caesium cathodes as illustrative examples is good, because they are widely used as well as typical, other cathodes perhaps deserve more than a brief mention, particularly the silver–oxygen–potassium and silver–oxygen–rubidium. For the experimenter, more complete information on the spectral response and thermionic emission of these cathodes, and of others, including those used for the ultra-violet, as well as on the spectral absorption of cell envelopes in ordinary and special glasses, would have been valuable. Also, much greater emphasis might have been laid on the importance of electrode design in cells to be used for purposes of measurement, so as to dispel any impression that one type of cathode is appreciably better than another in this respect. The differences in the metrological performance of different cells now on the market almost certainly result from differences of design, and not of cathode material. The production of a well designed cell is more difficult with some cathodes than with others, but the difficulties are not insuperable. To most scientific workers the photocell is essentially a measuring instrument, and one hopes therefore that manufacturers will not allow these difficulties to impede the regular production, in a form suited for precise work, of cells with a greater variety of cathodes including especially the newer very sensitive alloy cathodes. Furthermore, it is only with cells of such a kind that research on the photoelectric effect itself can give completely reliable results.

The monograph appears to be free from minor blemishes of consequence, though it is not quite true to state (p. 28) that no studies have been made on the spectral absorption of the Ag–O–Cs cathode film. Asao, in Japan, published the results of an interesting study of this subject in 1940.

The simplified theoretical introduction is good and quite adequate as a background to later chapters devoted to more detailed aspects. The subdivision under headings and sub-headings is clear and useful, and the style is simple and fluent.

Dr. Sommer has succeeded well in presenting his subject clearly and simply to the lay reader possessed of a certain background of general physical knowledge, while the amount of technical information compressed without loss of continuity into so small a volume, will undoubtedly commend the book also to a wider circle of users of photocells.

J. S. PRESTON.

A Handbook of Telecommunication (*Telephony and Telegraphy over wires*), by B. S. COHEN. Pp. xiv+437, with 281 diagrams and numerous tables. (London: Pitman and Sons, 1946.) 30*s*. net.

The arts of telephony and telegraphy, particularly the former, have made great strides in the past few decades. An account of the stages of this progress is of interest to the physicist as well as to the telecommunication engineer. This book provides a welcome medium for surveying the new territory. The author, B. S. Cohen, was well fitted to undertake the task of writing it. For many years he was in charge of the research work in the extensive laboratories of the British Post Office at Dollis Hill. He read several papers before the Institution of Electrical Engineers, and his name constantly appears in the discussions recorded in their Journal. He gave one of the Faraday lectures, his subject being " The Long-distance Telephone Call ". Before his premature death in 1940 he had fortunately completed the manuscript of this book. The final revision was undertaken by Mr. F. G. C. Baldwin with the assistance of some of Cohen's colleagues.

That the more striking advances lie in the branch of telephony is reflected in the preponderatingly large space the author has elected to give to it. After an introductory chapter, eleven chapters are devoted to telephony, one only to telegraphy, one to the thermionic valve, and one to measurements, the last two being more or less common to the two branches.

Speech and music can now be transmitted over wires almost without limit of distance over the earth's surface, with high fidelity of reproduction at the receiving end. In 1907 the highest indispensable frequency of reception was considered to lie between 800 and 1100 c./s. To-day it is internationally agreed that speech reception shall cover the band 300 to 3400 c./s, and broadcast music the band 50 to 6400 c./s. The improvement in quality is distributed over all stages, in the line and at each terminus. In the year quoted, Bell's receiver was much as it had left the inventor's hand in 1876. Today its frequency response curves are much nearer to the ideal, and the sensitivity is appreciably greater. In 1907, Pupin's loading coil was just coming into general service. In the quarter century that followed, several million pounds sterling had been invested in these coils in this country. By 1936 they were fast disappearing from the lines, to reappear in the terminal apparatus.

The introduction of the thermionic valve solved the problem of distant transmission. Black's 3-valve negative feed-back amplifier met the requirements in a highly effective manner. By its insertion at suitable intervals along the line the attenuated power is stepped up by 50 or more decibels at each stage, whilst close fidelity is preserved over the whole range of frequencies.

The great expansion of business over trunk lines raised acutely the problem of economical transmission. Here the adoption of the technique of radio is the remarkable feature, and the co-axial form of air-cored cable proves peculiarly appropriate to the purpose. Using, for instance, carrier-frequencies with an over-all range of 0·5 to 2·1 Mc./s., and allowing a band of 4000 c./s. for each message (only one side-band is transmitted, the carrier and the second side-band being suppressed) it is possible to send 400 independent messages simultaneously over a single pair of wires. The re-introduction of the carrier-frequency at the receiving end involves synchronization of the oscillator at that end to one part in two millions ! Band-pass filters play an essential part in finally resolving the messages.

In telegraphy, too, the carrier principle is employed to the same end, to provide the voice-frequency multi-channel telegraph system. Here the carrier frequencies lie within the audio range. The teleprinter is now in general use, and the vogue of the Morse code is rapidly becoming obsolescent.

These and other notable advances will be found described in some detail in Cohen's book. The task of selection of subject-matter amid the avalanche of change must have proved a delicate one. The reader has, however, been provided with copious references to text-books and original papers, a list of these being appended to each chapter. (The omission of Dr. Mallett's *Telegraphy and Telephony* from these lists is surprising).

A good index is especially valuable in a book of this kind. The index supplied, though running to 15 pp., is perhaps, hardly adequate. Our use of it has disclosed several omissions, and we would suggest that the index be considerably extended in a second edition.

The book, with its descriptions of the latest forms of microphones and telephone receivers and other apparatus, and its clear accounts of the application of physical principles in such important practical fields, will we think be warmly welcomed by teachers and students of physics, and should find a place in all scientific libraries. D. O.

Understanding Microwaves, by VICTOR J. YOUNG. Pp. 385 + xi. (New York: John F. Rider Publisher, Inc., 1946.) $6.00.

The first point to understand about any scientific or technical book is its title, and in this connection the reviewer has always been a little doubtful as to the meaning of the term " microwave " and the portion of the radio frequency spectrum to which it applies. In an earlier book entitled *Microwave Transmission*, by J. C. Slater, the first sentence of the preface states that " Microwaves are electromagnetic waves of wave-lengths that we may take, for definiteness, to be between 1 centimeter and 1 meter." In the absence of any specific information to the contrary we may accept this definition as applicable to the work under review, although it is perhaps doubtful if the author intended to be limited at the short end to a wave-length of 1 centimetre.

While it is not clear to what class of reader this book is addressed, it is to be noted that much of the latter portion of the book deals with microwave terminology and contains some relatively advanced mathematical conceptions which are in marked contrast to the elementary nature of the introductory chapters. In such a rapidly advancing subject as radio, it would seem unnecessary for modern books to recapitulate the elementary principles of electricity and magnetism, alternating current circuits and transmission lines, and yet the first five chapters of this book are confined to an exposition of these matters. Chapter 6 deals with Poynting's Vector and Maxwell's Equations; and here, as throughout the book, the mathematics is purposely reduced to a minimum, and the information is conveyed in a graphical and descriptive manner. On page 123 an incorrect formula is given for the complex dielectric constant of a conductor, which has the effect of making this constant directly proportional to the incident electric field instead of being independent thereof, and directly instead of inversely proportional to the frequency.

Some wave-guides and their properties are treated non-analytically in Chapter 7, and resonant cavities are described in Chapter 8, together with some applications. These two chapters give a reasonably clear picture of the basic principles involved, though the difficulty of presenting a concise and really adequate account without the use of mathematics is perhaps more apparent here than in other parts of the book. The following chapter contains a brief description of various types of centimetre wave antennae, including parabolic reflectors, horns and corner reflectors. In a short note on propagation at the end of this chapter reference is made to interference between direct rays and those reflected from the earth, in which it appears that the word interference is used to cover only those cases where the signal received is less than the free space signal. The term " interference ", however, covers the whole phenomenon and embraces both the " bright and dark " bands of the radiation field pattern. Chapter 10, entitled " Microwave Oscillators ", is probably the best chapter in the book. Although admittedly in keeping with the descriptive and rudimentary style of the rest, it nevertheless, within the limitations imposed by such a treatment, gives a clear picture of the manner of operation of klystrons and magnetrons, the valves which play such an important part in centimetre wave technique. The final chapter in Section I of the book gives an outline of the factors which limit the range and application of microwaves in radar and communications. Receiver sensitivity and bandwidth, antenna gain and attenuation during propagation are discussed in relation to this problem.

Section II, comprising the last 100 pages of the book, is devoted to a glossary of microwave terminology and represents a quite useful collection of definitions and theorems applicable in radio technique. It might be noted that (on page 297) the functions referred to should be Hankel and not Henkle. Rather than use the definitions given on pages 309 and 310 it is better to class Fraunhofer diffraction as that in which the radiation source and location of the pattern are each effectively at infinite distances from the diffracting obstacle, whilst Fresnel diffraction refers to the case of finite distances from the obstacle. " Line-of-Sight Range " is here defined as the maximum distance over which microwaves can be transmitted. Neglecting the case of anomalous super-refraction, this is hardly a good term to use under normal refraction conditions; for radio waves are refracted to a greater extent than light waves in the normal atmosphere, so that, in fact, although a radio receiver might be within the direct range of the transmitter, they need not be optically intervisible.

The author is well aware of the difficulties of the units question and uses in the main the Gaussian system, although occasionally other systems, notably the M.K.S., are applied. It might have been preferable to have used a uniform system throughout.

In general, the book appears to contain a large amount of unnecessary material, while at the same time insufficient space is devoted to matters peculiar to very short wave technique; the portions dealing with antennae and with the propagation of waves could have been expanded with advantage.

R. L. S.-R.

/*A First Course in Mathematical Statistics*, by C. E. WEATHERBURN. Pp. xv+271. (Cambridge: The University Press, 1946.) 15*s*.

This book, one of the very few which treat of statistical theory in the form which it has attained in the last 30 years, is very definitely a companion to Fisher's *Statistical Methods for Research Workers*. The latter gives, without many concessions to the weaker brethren, a critical statement of the methods to be applied in testing significance and in analysing experimental results. Professor Weatherburn's book, on the other hand, gives the basic theory leading to the methods described by Fisher—theory which for the most part has had to be read in the original papers of " Student ", Fisher and other workers. It is true that a few good text-books have appeared in recent years in which the theory is accurately set out, but none of them could be described as suitable for the beginner.

Professor Weatherburn has therefore rendered a distinct service by writing this book, in which the reader is led as far as an introduction to the analysis of variance, and to the topic of multiple correlation. Before this, the reader has been introduced to small-sample theory and to the χ^2 test of goodness of fit. Much of this involves proofs of the distributions to which various types of statistic tend, and hence has been preceded by descriptions of different standard distributions, the whole introduced by two chapters in which the general notion of a frequency distribution and its chief parameters has been introduced, and the relation of a probability distribution to a frequency distribution explained. From a fundamental point of view this is, perhaps, the most important step in the whole subject. It is here treated well, though very concisely.

From this summary it might appear that, with the exception of chapters 1 and 2, the book merely duplicates Fisher's work, but there is in fact a great difference, and the two are, as the author claims, complementary. Weatherburn gives the theory (in elementary form) and illustrates it with examples in which pains have been taken to keep the actual arithmetical work within bounds. The book does indeed form a good introduction to the theory, and introduces methods, like the moment-generating function and the cumulative functions, which Fisher, in his text-book, has no need to mention. In addition to exercises for practice, each chapter has a set of references for further reading, which the student is advised to peruse while he is working through the book. These should be of considerable value—there is a great deal to be gained by reading the same thing in different words or in a different order—though at first sight the lists seem rather intimidating, owing to their length.

The book is beautifully produced, and seems remarkably free from misprints.

J. H. A.

THE PROCEEDINGS OF THE PHYSICAL SOCIETY

VOL. 59, PART 2 1 March 1947 No. 332

A NOTE ON THE INTERMITTENCY EFFECT

By R. WEIL,
South-West Essex Technical College

Communicated by Dr. H. Lowery ; *MS. received* 10 *July* 1946

ABSTRACT. The nature of the intermittency effect is described and a new criterion for its detection mentioned. This is applied to the manifestation of the effect when an alternating current light source is used. The experimental arrangement employed in the present investigation is given, and the results of the latter are briefly discussed.

§1. INTRODUCTION

NUMEROUS workers have found that photographic materials are, in general, unreliable in integrating an intermittent exposure. Thus the density produced on a plate by ten one-tenth-of-a-second exposures will differ from that due to a single exposure of the same intensity lasting one second. This fact may be found when both the intermittent and the continuous exposures are made on one plate: it is due to an objective phenomenon resulting from the way in which light reaches the individual grains of the emulsion, and is known as the *Intermittency Effect.*

In connection with intermittent exposures it has also been found that the Bunsen-Roscoe reciprocity law breaks down. According to the latter, $I \times t =$ constant, where I is the intensity of the light and t is the time of exposure. This law applies only to the straight-line portion of the curve connecting the density d and the logarithm of the exposure. Its breakdown implies that the ratio of two densities does not equal the ratio of the logarithms of the corresponding exposures (exposure $E = I \times t$).

The intermittency of illumination is generally produced by means of a rotating sector-wheel. Talbot was the first to suggest that if such a wheel were rotated at a suitably high velocity the eye would obtain the impression that any light behind the wheel was of intensity I', whereas its true intensity was I. The ratio I'/I would be given by

$$I'/I = \theta/2\pi, \qquad \ldots\ldots(1)$$

where θ is the angle of the sector.

In a series of brilliant researches, Webb demonstrated the true nature of the intermittency effect. He found that it was closely connected with the failure of the Bunsen-Roscoe reciprocity law. It should be pointed out that, because of these effects, it has become necessary to introduce two scales in photographic measurements: the intensity-scale in which the intensity is variable and the time-factor constant, and *vice versa* for the time-scale (figure 1). Webb found that the intermittent time-scale curve, which lies generally between the two others, coincides with the former if the sector-wheel is run above a critical speed. (Webb and Silberstein (1934) have later shown that such a coincidence is not possible theoretically but may be considered to exist for all practical purposes.) If light is propagated in discrete quanta, it is evident that any exposure is really intermittent. On the basis of this conception, Webb showed that the critical speed mentioned above was such that, on the average, not more than one quantum of light per flash falls on one grain of the photographic emulsion. The critical speed was given by the relation

$$f_c = g \,.\, I_{\mathrm{av}}, \qquad \ldots\ldots(2)$$

where g is the effective reception area of the grain and I_{av} the average intensity.

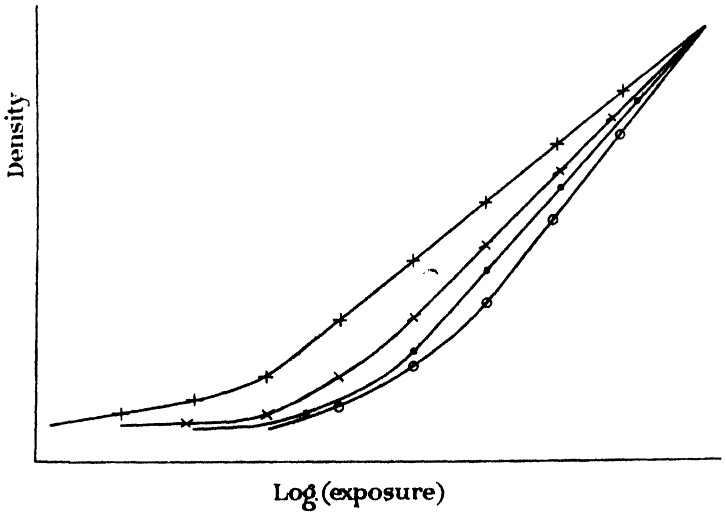


Figure 1. + Time scale. × Low frequency of flash. ● High frequency of flash. O Intensity scale.

This equation states that the critical frequency is proportional to the average intensity. It is plain that the probability of one quantum being transmitted by the sector-wheel is greater for a given frequency of flash if the sector-aperture is larger. It is therefore to be expected that if two different sectors are run at the same speed (cf. Lochte-Holtgreven and Maecker (1937)), the intermittency effect will or will not be apparent according as the speed is below or above the critical speed for the larger sector. Since rotating sector-wheels are used in measurements of reflectivity and absorptivity, and, therefore, must have variable apertures, it is vital that the above point should be borne in mind.

§ 2. EXPERIMENTAL ARRANGEMENT

In order to investigate the possibility of using A.C. sources for photometry by photographic methods, the author designed a special sector-wheel (figure 2). As has been mentioned before, rotating sectors are used in measurements of absorption and reflectivity: the beam which is not absorbed or reflected is passed through the wheel and the sector of the latter adjusted until a match is obtained between its intensity and that of the other beam. When the speed of the wheel is very high it is necessary to ensure that it should be as symmetrical as possible to prevent it from breaking owing to centrifugal forces. That is why two sectors facing each other were cut into the wheel.

It has also been shown that if both beams are interrupted at the same rate, the intermittency effect will be eliminated provided that the frequency of flash corresponds to, or is higher than, the critical frequency for the larger aperture. Lochte-Holtgreven and Maecker, who used a rotating sector-wheel in their work on the temperature of a freely burning carbon arc, employed a dummy sector in the path of the standard beam to eliminate the effect. In the present arrangement only one wheel is used, and both beams are passed through it: the beam to be reflected or absorbed through the large angle, which is fixed (at 80°), and the other through the variable sector. The latter has a radius only one half that of the wheel, and therefore cannot interfere with the passage of the beam through the large sector. For a given intensity there will be a marked intermittency effect at low frequencies of flash. But for every flash the current will alternate many times and, on the average, the stroboscopic effect will be negligible. As the speed of rotation is increased, the intermittency effect proper will be eliminated, but there will be fewer current-cycles per flash and a secondary effect would become apparent. In fact, it was seen that when the sector-speed was 32 r.p.s.—a speed at which no individual flashes can be perceived—a distinct flicker was observed when the A.C. source was viewed: sunlight, however, reflected from a mirror, and transmitted

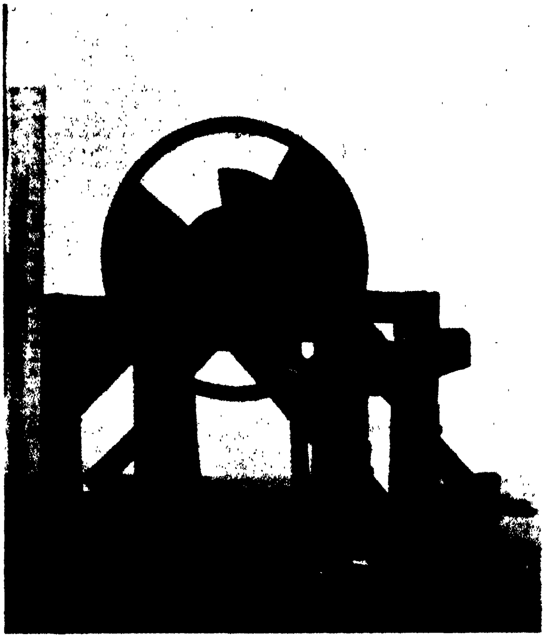


Figure 2.

through the sector-wheel running at the same speed, appeared to be perfectly steady. This secondary effect would be expected to affect the photographic plate in exactly the same way as the ordinary effect in connection with D.C. sources of light.

It is seen from figure 1 that the intermittency effect gives rise to a density difference for equal exposures owing to the existence of different curves for the two beams if only one is interrupted. Thus, if the difference in density is plotted against the corresponding difference in the logarithms of the exposures for one standard beam, curves will be obtained which do not pass through the origin if the intermittency effect is present. Conversely, if they pass through the origin its elimination may be assumed. In particular, if the straight-line portions of both curves are used—a condition which can be fulfilled only with plates having a small γ—straight lines will result whose tangent is equal to γ, but which will pass through the origin only if the intermittency effect is eliminated. This will be the case when the exposures due to both beams lie on the same curve. Although the theory does not apply to heavily over-exposed plates, it is evident from figure 1 that in the case of too long or too short exposures, all the curves coincide and no conclusive results can be obtained. This fact is only mentioned because it is not essential that our measurements be confined to the straight-line portion of the curve, although this is desirable. Indeed, as can be seen from figure 3, with very contrasty plates it is practically impossible to work on the straight-line portion only.

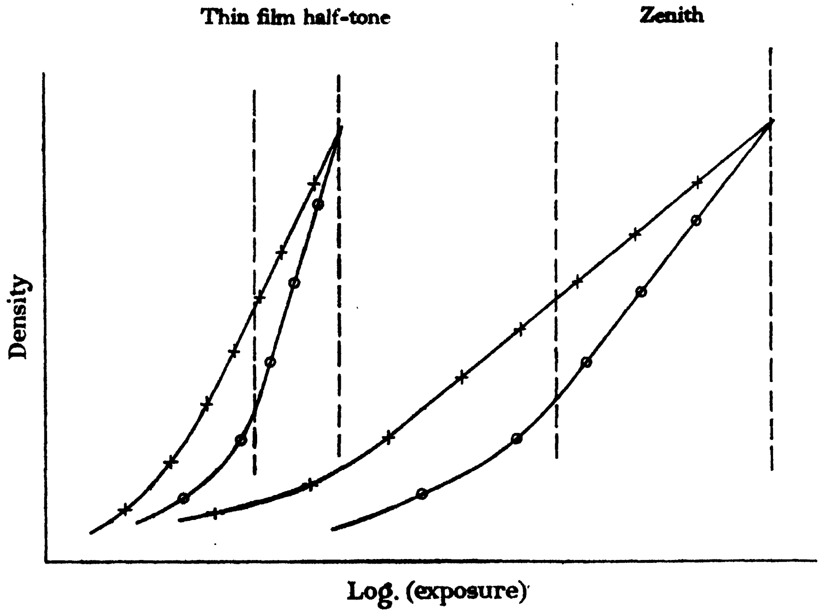


Figure 3. Range of straight line portion.

Since the exposure times are proportional to the two sector apertures, the densities produced by the two parts of the wheel will be given by the expressions:

$$d_1 = k \cdot \log I\alpha_1 + b, \qquad \ldots\ldots(3)$$

$$d_2 = k \cdot \log I\alpha_2 + b, \qquad \ldots\ldots(4)$$

where α_1 and α_2 represent the angular apertures of the sectors and b is a factor representing any accidental fogging. Since the two blackened areas on the plate are near to each other, b may be assumed to be the same for each of them. Subtracting the second equation from the first,

$$d_1 - d_2 = k \cdot \log \frac{\alpha_1}{\alpha_2}. \qquad \ldots\ldots(5)$$

This expression refers only to the straight-line portion of the curve. But if an arbitrary point is taken on the intensity-scale curve in figure 1, it will be obvious that, although no straight-line graphs need necessarily be obtained, the curve will pass through the origin if the intermittency effect is absent.

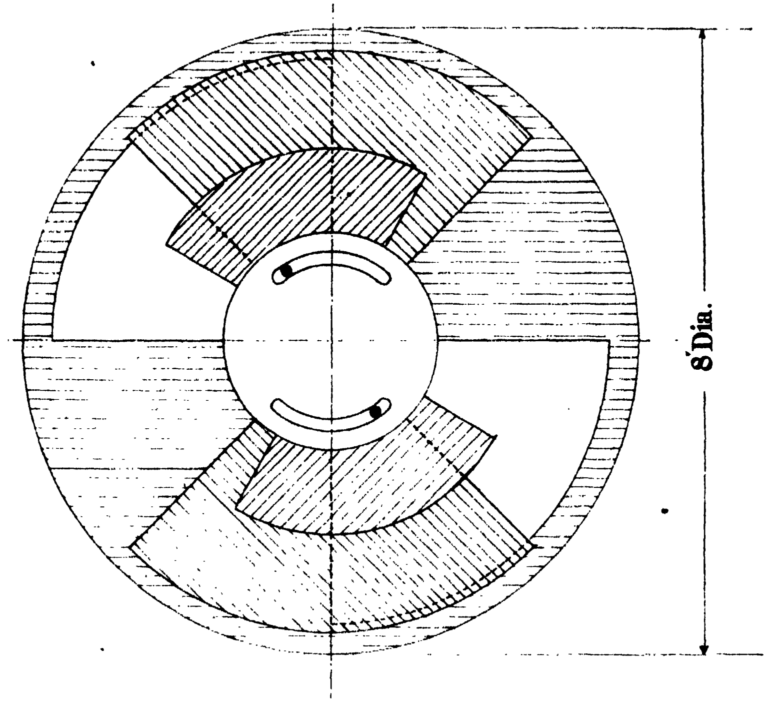


Figure 4.

The sector-wheel consisted of a brass plate 8" in diameter and $\frac{1}{16}$" thick, with two sectors cut into it diagonally opposite each other (figure 4). A large vane allowed the sector-apertures to be reduced from 90° to 0°: during the subsequent investigation the apertures were kept at 80°, as this facilitated the procedure. Another vane, similar to the previous but with only half the diameter, permitted the further cutting down of the central parts of the apertures. The whole arrangement was painted black. A ground-glass screen illuminated with the diffused light of a filament lamp was placed immediately in front of the wheel (figure 5). A lens focused the screen on the photographic plate, and a green filter (Ilford 404) was interposed. It was desirable to use monochromatic light since, according to Webb, the critical frequency varied with the wave-length of light. The sector-wheel was driven by an electric motor whose speed could be varied. The two turned at the same speed, which was measured on the shaft of the motor by means of a tachometer.

Before the actual experiment was embarked upon, the characteristic curves of the two plates in use were determined so that extremes of exposure could be avoided.

As mentioned above, the aperture α_1 was kept at 80° throughout the experiment. On any one plate, for which α_2 was kept constant, eight photographs were taken for frequencies of flash ranging from 4 to 32 per second. Since only the difference in densities is considered, it would appear to be in order to combine the results for different samples of one kind of plate. Various objections might be raised in connection with this procedure, but it is difficult to justify an alternative. The object of the investigation was to find variations which depended on the frequency of flash. Evidently there was a better chance of discovering them in the above manner than if different sector-apertures had been photographed on one plate and a different speed of rotation chosen for every plate. A test for the reliability of the procedure was afforded by the regularity of the densities of the patches blackened by a continuous exposure, as will be seen later. In order to demonstrate the justification of the use of a single wheel for the two beams, a series of photographs was taken in which only one beam was interrupted, the other subjecting the plate to a continuous exposure. The ratio α_1/α_2 was taken to be $2\pi/\theta$. Two plates were used: Zenith (H & D 700) and Ilford Thin Film Half-Tone. Their relative slopes are indicated in figure 3, and it is evident that the range for work along the straight-line portion is very small in the latter case. The times of exposure varied from 40 to 100 seconds, but were kept constant for any particular series of exposures. The usual development technique was used. In order to reduce any error due to the use of different samples of one kind of plate, the times of development were chosen so as to obtain a maximum γ, in fact γ_∞. That this has been achieved has already been stressed above.

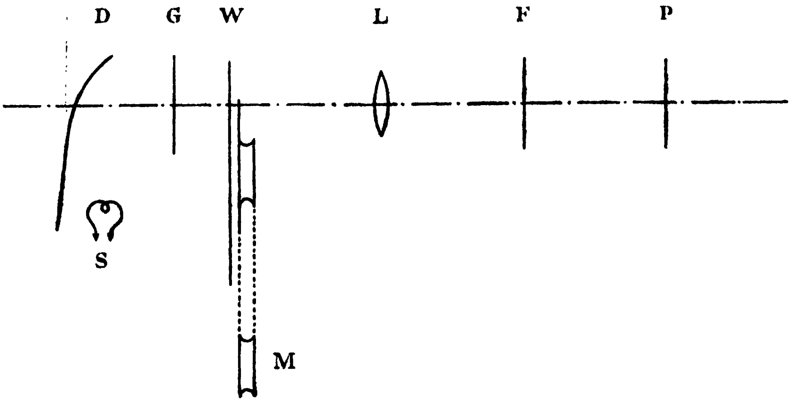


Figure 5.

S.	Light source.	M.	Electric motor.
D.	Diffusing screen.	L.	Lens.
G.	Ground-glass screen.	F.	Green filter.
W	Sector-wheel.	P.	Photographic plate.

The plates were subsequently examined with a microphotometer and graphs plotted as suggested above.

§ 3. RESULTS

In the cases when only one part of the exposure was intermittent (and the other continuous) irregular results were obtained for the Zenith plates; those for the Thin Film are shown in figure 6. It is clear that a density difference exists for the lower frequencies of flash and that, at higher frequencies, the curves become irregular owing to secondary effects which are connected with the alternating source of light.

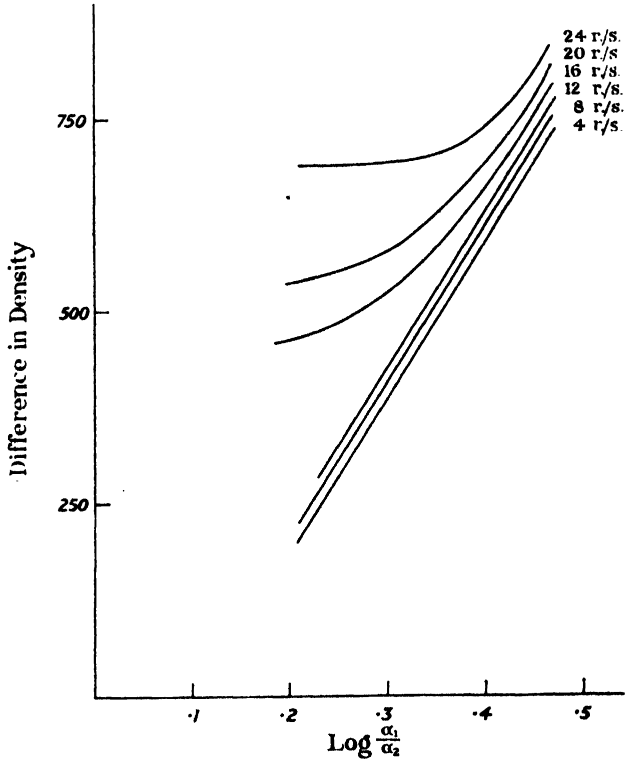


Figure 6. Thin film half-tone (single sector).

But with both beams interrupted at the same rate, there is an intermittency effect at the lower frequencies only, and this disappears as the speed of rotation is raised (figures 7 *a*, 7 *b*). This is in agreement with expectations. The fact that the critical frequencies for both kinds of plate lie in the neighbourhood of 12 r.p.s. is probably accidental as a different lamp was used for each kind. At frequencies of 28 flashes per second and more, the intermittency effect reappeared. The curves representing 28 and 32 r.p.s. are not indicated in the figures as they would tend to obscure rather than clarify the point in question. Let it be said, however, that they do not pass through the origin but approach the positions of the curves for less than 16 flashes per second. The stroboscopic effect, so obvious to the eye, is also affecting the photographic plate.

§4. CONCLUSION

Webb's work has shown that the intermittency effect can be eliminated when D.C. sources of light are used if the frequency of flash is above a certain critical value. From the above work it is seen that the same applies to A.C. sources on these conditions: (*a*) the standard beam and the beam to be reflected or absorbed must be interrupted at the same rate; (*b*) the frequency of flash must be greater than a certain critical frequency; (*c*) the frequency due to the stroboscopic effect, i.e. the combination of the frequencies of the A.C. source and the sector wheel, must be above the same critical frequency. This latter statement was substantiated by visual observations.

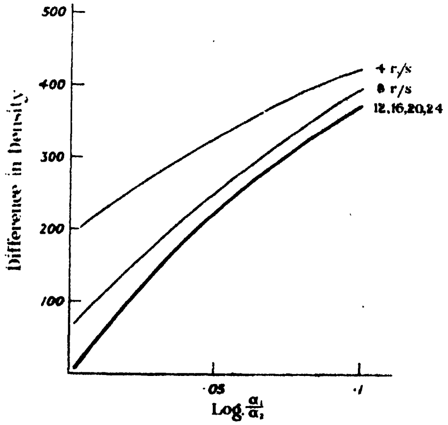


Figure 7 *a*. Thin film half-tone (double sector).

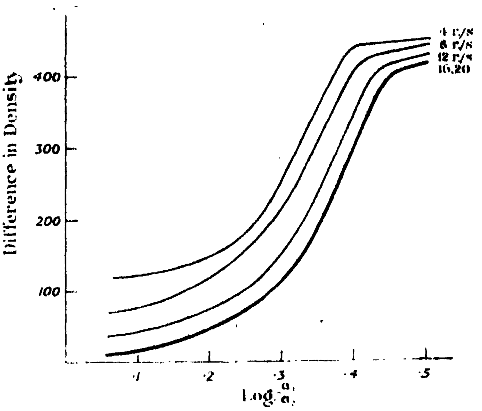


Figure 7 *b*. Zenith (double sector).

§5. ACKNOWLEDGMENTS

The above investigation was carried out under the auspices of the Pyrometry Sub-Committee, British Iron and Steel Research Association.

The writer would like to thank Dr. H. Lowery, South-West Essex Technical College, for his interest and encouragement, and Mr. G. Lothian, M.A., for many valuable discussions. Sir Lawrence Bragg, F.R.S., Cavendish Laboratory, Cambridge, kindly made available a microphotometer. The work would have been rendered very cumbersome without the good offices of Miss M. MacIntyre and Mr. W. Bennett.

REFERENCES

FORSYTHE, 1940. *Measurement of Radiant Energy* (McGraw Hill), Chapter VIII *.
LOCHTE-HOLTGREVEN and MAECKER, 1937. *Z. Phys.*, **105**, 1.
SILBERSTEIN and WEBB, 1934. *Phil. Mag.*, **18**, 1.

* This chapter contains a very extensive and up-to-date bibliography on the intermittency effect.

DELAYED FRACTURE IN GLASS

By C. GURNEY,
Royal Aircraft Establishment, Farnborough

MS. received 6 *August* 1946

ABSTRACT. An important delayed-fracture effect is found for mineral glass. A likely cause might be the gradual spread of Griffith cracks, but a theorem due to Griffith states that cracks do not spread below a certain average stress, and that at this stress they spread catastrophically. In the present paper, Griffith's theorem is re-examined and it is shown that in materials having atomic constitution, fracture does not occur catastrophically at the Griffith stress ; the Griffith stress is the least at which a crack may start to spread by a process of splitting, and the rate of spread is controlled by the rate at which thermal motions overcome energy barriers. Catastrophic fracture does not occur until the stress at the end of the crack equals the maximum a material can withstand. An improved method of estimating this stress from thermal data is given. The best estimate is that the maximum stress is equal to the intensity of the given stress system which makes the latent heat of evaporation zero. The rate of spread of cracks by Griffith's process is considered to be too slow to account for much of the delayed-fracture effect in glass. Other processes are considered under the headings of approach to homogeneous and heterogeneous equilibrium. The attainment of homogeneous equilibrium under stress in materials in equilibrium when stress-free involves, in the absence of phase changes, but a small entropy change, and is not likely to cause an appreciable time effect. Glass, however, is not in thermal equilibrium when stress-free, the high temperature phase persisting at temperatures below the transition point. On account of the high internal viscosity of glass, approach to equilibrium effectively ceases soon after manufacture. Stress reduces the internal viscosity and enables approach to equilibrium to continue. Because of the concentration of stress at the ends of the cracks, the approach to equilibrium made possible by stress is much faster in the material at the ends of the cracks than elsewhere, and as attainment of equilibrium involves volume shrinkage, the stress at the end of the crack is increased and the crack spreads. This effect would be expected to cause an appreciable delayed fracture effect.

Two effects are considered under the heading of approach to heterogeneous equilibrium. The first is evaporation of the material at the end of the crack. An estimate of delayed fracture due to this cause suggests that it is unimportant. A much more important cause of delayed fracture is atmospheric attack of the glass. Due to concentration of strain energy, the material at the end of the crack has a much higher free energy than normal unstressed glass, and is therefore much more chemically active. Atmospheric attack will result in the formation of a complex of glass and atmospheric constituents. The crack will extend continually if the strength of this complex, during or after its formation, is less than the load imposed on it.

Changes due to stress in the stable phase at room temperature are considered, but these are not likely to be important for glass, as in this material the high-temperature phase persists at room temperature. Phase changes under stress may have important effects on the behaviour of other materials. Included in the paper are formulae for the stress coefficients of vapour pressure, entropy and phase-transition temperature, of material subjected to a generalized stress system.

§1. INTRODUCTION

A VERY remarkable delayed-fracture effect is found in mineral glass, the time to fracture increasing with decreasing load, the earliest investigation being due to Grenet (1899). More recently Preston (1942),* quoting results obtained by T. C. Baker, found that the time for annealed soda lime glass to

---

* Since this paper was written a fuller account of the work of Preston and his collaborators has become available. It is given in papers by Preston, Baker and Glathart in *J. Appl. Phys.*, **17**, 162 (1946).

fracture in bending increased from 10^{-2} sec. to 10^4 sec. when the load was reduced in the ratio 3 : 1. Somewhat similar results were obtained by Holland and Turner (1940). According to Griffith (1920), whose theory has been accepted by many subsequent investigators, glass contains submicroscopic cracks, and its comparative weakness is due to concentration of stress at the ends of these cracks. A likely explanation of the delayed fracture of glass would be the gradual growth of these cracks in material under load. Griffith, however, considered the conditions for crack growth, and decided that, below a critical load, crack growth could not occur, and that at this load the crack spread catastrophically. The critical load was that at which the spreading of the crack would result in a reduction of the free energy of the system.

Recent attempts at the explanation of the delayed fracture of glass have been consistent with Griffith's theorem. Orowan (1944) has suggested that the main effect may be caused by atmospheric attack of the surface of the glass. He suggests that the surface tension of contaminated glass may be about one-tenth that of a freshly formed glass surface. Griffith's energy theorem gives the stress to fracture to be proportional to the square root of the surface tension, so that contamination of the surface reduces the stress at which the crack spreads in the ratio of about 3 : 1. When the load applied to glass is sufficient to break it quickly there may be no time for contamination of the rapidly formed surfaces of the spreading crack. In this case, a stress corresponding to the higher surface tension of the glass is obtained. If the load is not sufficient to break the glass quickly, but exceeds that necessary to cause a crack to spread if the surface tension has its lower value, the crack will gradually spread. The rate of spreading will depend on the time for the contamination of the freshly formed surface to occur to an extent sufficient to reduce the surface tension to that necessary for the crack to spread.

Orowan's explanation of the delayed-fracture effect implies that the strength of glass as measured in an ordinary quick-loading test will be reduced by preloading the glass at a load sufficient ultimately to break it. Murgatroyd and Sykes (1945), who experimented with preloaded test pieces, found no significant reduction in strength, and Murgatroyd has proposed an alternative explanation of the delayed-fracture effect. Murgatroyd (1944) has suggested that the Griffith cracks are filled with a viscous constituent of the glass which gradually relaxes under load. At the commencement of loading there is no concentration of stress at the ends of the cracks, but as the viscous material relaxes, concentration of stress at the ends of the cracks increases until fracture occurs. The present author in collaboration with Pearson (1946) has made experiments on the delayed fracture of round glass rods under four-point bending. Three series of tests were made. In the first the test pieces were not rotated and in the second and third the test pieces were rotated at 14 and 10,000 r.p.m. respectively, the direction of loading being stationary. The curves relating stress and time to fracture for the three conditions did not differ significantly. This result is at variance with Murgatroyd's theory. Under cyclic stress, the viscous constituent of Murgatroyd's model would not be expected to relax continuously, and delayed fracture under cyclic stress would not be expected.

It seemed to the present author that Orowan's and Murgatroyd's explanations were restricted by their acceptance of the Griffith load as being that at which cracks spread catastrophically, and below which cracks did not spread at all. In this paper Griffith's theory is reviewed, and it is concluded that it is consistent with the gradual spread of cracks even in the absence of atmospheric attack. Catastrophic fracture does not occur until the stress at the end of the crack reaches the maximum the material can withstand. § 2 of this paper therefore considers the estimation of this stress. In § 3, crack spreading by splitting in the manner envisaged by Griffith is treated. In §§ 4 and 5 other processes leading to delayed fracture are suggested. These result from the stressed material approaching thermal equilibrium. It has been found convenient to distinguish effects related to homogeneous equilibrium from those related to heterogeneous equilibrium. Homogeneous equilibrium is concerned with the arrangement of atoms in a given phase. As the strength of material depends on the atomic arrangement, gradual rearrangement of atoms will result in a gradual change of strength. Heterogeneous equilibrium is concerned with the partition of material amongst the various phases, as for example the number of atoms in the vapour phase in a material in equilibrium with its vapour; evaporation of material at the ends of cracks would lead to cracks spreading and delayed fracture. In § 6, phase changes caused by stress are considered.

Before proceeding, the application of thermodynamics to the minute quantity of highly stressed material at the bottom of the cracks needs consideration. The laws of thermodynamics are statistical laws, and strictly apply only to assemblies containing large numbers of elementary units. As the number of units is reduced, the variations of the results of individual experiments from the mean result increases, the standard error varying inversely with the square root of the number of units, but no change in quality of the laws appears until very few repetitions of experiments involving very few atoms are considered. In the latter case the laws cease to hold; for example, in a crystalline material, the jump of an atom from a lattice point to an interstitial position can occur although it causes an increase in free energy. In the case of cracks spreading, however, movements of quite large numbers of atoms are involved. Griffith cracks in glass are estimated to be of the order of 10^{-3} cm. long. Stress is proportional to the square root of crack length, so that to account for delayed fracture at 70% of the quick-loading strength, the cracks must double their length. If only the single string of atoms along the longest line contour of the crack is affected, the number of atoms involved in doubling the length of the crack will be of the order of 10^{10}, so that statistical laws should apply without very great variation in results.

Another subject which needs a brief discussion is the application of thermodynamics to stress systems other than hydrostatic pressure. Formulae for the stress gradient of vapour pressure of material under a generalized stress system will be given later. In general, the vapour pressure differs on each of the three pairs of parallel faces which bound a cube of material, and the system therefore cannot be in equilibrium. If, however, the time for any effective change to be caused by evaporation is long compared with the time for some other change, e.g. chemical combination, to occur, the condition can be regarded as one of metastability, and thermodynamic formulae can be applied. It may be worth mentioning here that the free energy of any thermodynamic system involving other than hydrostatic pressure may be reduced by the material spontaneously separating or shearing along internal surfaces; because the work done by the external forces exceeds the surface energy of the fracture surfaces. Usually, however, the time for spontaneous fracture is so long compared with the time for fracture by crack spreading, that spontaneous fracture is not a factor of any practical importance.

§2. THE STRENGTH OF FLAWLESS MATERIAL

The strength of flawless materials is usually estimated from thermal data such as latent heat, surface tension and triple point or from a theory of atomic binding. A summary of this work is given by Houwink (1937). The estimate of strength given here does not differ greatly in principle from previous estimates, but it gives them more precise expression.

A thermodynamic system under uniform pressure becomes unstable when the partial derivative of the pressure with respect to the volume becomes positive. By similar reasoning, material under a generalized stress system will become unstable when the partial derivative of the load in one direction with respect to the movement in that direction becomes negative, load and movement being of the same sign when in the same direction. Instability in a liquid under hydrostatic pressure results in the sudden formation of two phases. In a solid under generalized stress, instability with respect to a tensile stress results in fracture. In terms of atomic constitution, fracture would be expected to occur virtually immediately on loading, if the average thermal motion was just sufficient to enable atoms associated with the average strain energy to reach the top of the average energy barrier. If more than the average thermal motion is necessary, fracture may still occur, but it will not be immediate. Individual atoms will occasionally acquire sufficient kinetic energy to enable them to migrate, leaving a hole which by the same process will gradually develop into a crack and lead to fracture.

Physical knowledge of glass is not at present adequate for the above criterion to be used to estimate the maximum strength of flawless material. It is however possible to estimate the stress at which the average thermal motion is sufficient to vaporize the material, for this is the stress at which the latent heat of isothermal evaporation is zero. This stress may be expected to exceed the true breaking stress by a factor of two or more, but it seems the best estimate of theoretical strength at present possible.

The calculation of the stress at which the latent heat of evaporation is zero may be made by formal thermodynamic methods. It is first necessary to calculate the change of vapour pressure with stress. A three-dimensional stress system can be denoted by X_{ij}, where i and j take values 1, 2, 3; the corresponding strain system can be denoted by e_{ij}. The differential of strain energy per unit volume is then

$$\Sigma X_{ij} de_{ij}. \qquad \ldots\ldots(1)$$

This expression is the change of internal energy due to external work done on the material. It replaces the more familiar *pdv*, which applies to the hydrostatic pressure case. The calculation proceeds by stating that at equilibrium a small change in the system must be reversible, and that in a reversible isothermal process, change in the Helmholtz free energy is equal to the external work done on the system. It is found that the vapour pressure varies with the direction of the surface from which evaporation is considered to take place. If any particular stress is denoted by X_{kl}, the partial change in vapour pressure (p_{mm}) of the surface whose normal is mm for $kl \neq mm$ is given by

$$\frac{\partial p_{mm}}{\partial X_{kl}} = \frac{\Sigma_{ij} X_{ij} \frac{\partial e_{ij}}{\partial X_{kl}} - X_{mm} \frac{\partial}{\partial X_{kl}} \Sigma_i e_{ii}}{\frac{V^G}{V^S} - 1 + \Sigma X_{ij} \frac{\partial e_{ij}}{\partial X_{mm}} - X_{mm} \frac{\partial}{\partial X_{mm}} \Sigma_i e_{ii}}. \qquad \ldots\ldots(2a)$$

If $kl = mm$, we have

$$\frac{\partial p_{mm}}{\partial X_{mm}} = \frac{\Sigma_{ij} X_{ij} \frac{\partial e_{ij}}{\partial X_{mm}} - X_{mm} \frac{\partial}{\partial X_{mm}} \Sigma_i e_{ii} - 1}{\frac{V^G}{V^S}}. \qquad \ldots\ldots(2b)$$

Here V^G and V^S are the specific volumes of gas and solid respectively.

The above expressions give the partial change of vapour pressure of the surface whose normal is the direction mm, when surfaces normal to other axes are not acted on by vapour pressure. To evaluate these expressions accurately it is necessary to know the elastic constants and equations of state of gas and solid over the whole stress range. In the absence of this knowledge the best that can be done is to assume the elastic constants invariable, and to use the values obtained by experiment at low stresses and pressures. For the simple case of unidirectional tension acting on an isotropic solid, and considering evaporation off a stress-free face, these expressions give, for a vapour which is a perfect gas, and whose pressure is small compared with X,

$$\ln \frac{p}{p_0} \approx \frac{V^S X^2}{2ERT}, \qquad \ldots\ldots(3)$$

where p and p_0 are the vapour pressures at stress X and at zero stress respectively, and E is Young's modulus. If a value of the stress-free vapour pressure and of the latent heat of evaporation are known at a high temperature, those at any other temperature may be estimated by using the Clausius-Clapeyron equation and the equation giving the temperature gradient of latent heat. The latent heat of a reversible change is the product of the temperature and the change in entropy on passing from one state to another. The stress coefficient of entropy can be expressed in terms of derivatives of stress and strain by making use of the fact that some thermodynamic functions are total differentials, and that for such the order of differentiation is irrelevent. The process for non-hydrostatic systems is formally similar to that for hydrostatic stress. The result is

$$\frac{\partial S}{\partial X_{kl}} = V\left[\frac{\partial e_{kl}}{\partial T} + \left(\Sigma X_{ij} \frac{\partial e_{ij}}{\partial T}\right)\left(\Sigma \frac{\partial e_{ii}}{\partial X_{kl}}\right) - \left(\Sigma \frac{\partial e_{ii}}{\partial T}\right)\left(\Sigma X_{ij} \frac{\partial e_{ij}}{\partial X_{kl}}\right)\right]. \qquad \ldots\ldots(4)$$

As the stress varies, the equilibrium vapour pressure varies, so that for the solid

$$\frac{\partial S}{\partial X_{kl\,(p \text{ varying})}} = \frac{\partial S}{\partial X_{kl\,(p \text{ constant})}} + \frac{\partial S}{\partial p} \frac{\partial p}{\partial X_{kl}}. \qquad \ldots\ldots(5)$$

For the gas, the entropy change is simply

$$\frac{\partial S}{\partial X_{kl}} = \frac{\partial S}{\partial p}\frac{\partial p}{\partial X_{kl}}. \qquad \ldots\ldots(6)$$

For any particular stress system, equations (2), (5) and (6) can be solved by graphical methods and the intensity of stress at which L is zero, i.e. the stress at which S is the same for vapour and gas, can be estimated. In the absence of more precise information, Hooke's law and constancy of coefficient of expansion can be assumed. When making the calculation for silica glass in simple tension, it was found that the specific volume of the vapour at high stress was so small that the assumption of a perfect gas was clearly inadmissible; but as the stress varies as the square root of the logarithm of the vapour pressure, the stress calculated on this assumption should not be greatly in error. For practical purposes the maximum stress is reached when the vapour pressure is so high that the solid rapidly disappears by evaporation. For silica glass the maximum stress is estimated to be about $\pm 8 \times 10^6$ lb./sq. in. It has already been mentioned that the stress at which the latent heat of evaporation is zero exceeds the breaking stress by a factor of the order of two, and the assumption of invariance of the various coefficients leads to further error. The figure obtained corresponds to a breaking strain of 80%, and it is over twice the strength of fine silica fibres obtained experimentally by Anderegg (1939).

§3. GRIFFITH'S CRITERION AND DELAYED FRACTURE

Griffith considered the equilibrium, when stressed, of a brittle material containing cracks. Consider a rod of such material placed vertically, supported at its upper end, and supporting a load at its lower end. If an initial crack in the material were to spread a little, the load would fall a little, the strain energy of the material would increase slightly, and the total surface energy of the crack surface would increase, due to the increase in area. The loss of potential energy of the load and the gain of strain energy for a given increment in crack length are proportional to the length of the crack, while the gain in surface energy is independent of crack length. For very small crack lengths, the gain in surface energy is greater than the difference between the loss of potential energy and the gain in strain energy, and the crack therefore does not tend to spread. Above a certain crack length the reverse is true, and at a critical crack length, for a given load, the difference between the loss of potential energy and the gain in strain energy produced by an increment in crack length just equals the gain in surface energy. Griffith (1920) considered that at this crack length failure would be immediate, because "the system can pass from the unbroken to the broken condition by a process involving a continuous decrease in potential energy". In a given material with cracks of given length the criterion gives the load at which fracture occurs. Griffith therefore considered that fracture is immediate if spreading of the crack decreases the free energy of the system (all the energies considered by him are free energies). From the quotations from his paper it is clear that he is considering materials to be composed of continuous elastic media, and, applied to such, his critical load would produce immediate fracture. The position is different however when the atomic constitution of materials is introduced. For such materials, even though a state of lower free energy exists, the transition to this state takes time on account of the energy barriers which have to be overcome. For example, the free energy of the system comprising a rod in tension supporting a weight would be reduced if the rod flowed and increased its length while reducing its cross-sectional area, but it is a matter of common experience that, in many materials, the rate of flow is so slow as to be negligible at room temperature. This is because the energy of the average thermal motion is small compared with that of the average energy barrier, and only very rarely does the kinetic energy of atoms exceed that of the average by a quantity sufficient to overcome energy barriers. Other examples of similar delay in attainment of equilibrium are the time taken for materials surrounded by an unsaturated atmosphere to evaporate, and the time for viscous materials, such as glass, to crystallize when maintained at temperatures below the freezing point. The reduction of free energy as a crack spreads is a necessary condition for its spreading, but, from the examples quoted, it is clear that this is not the condition for the crack to spread catastrophically. The rate of spreading will depend on the rate at which thermal motions overcome energy barriers; if much more than the average thermal motion is necessary, the rate of crack-spreading may be negligible. The crack will spread catastrophically only when the average thermal motion is sufficient to overcome the average energy barrier, that is when the stress at the end of the crack equals the maximum the material can withstand.

A possible explanation of the decay in strength of glass is, therefore, the gradual spread of cracks by the material splitting in the way suggested by Griffith.

The thermal energy of mineral glass at room temperature is of the order of a few per cent of the latent heat of vaporization at room temperature, and, therefore, is probably not more than ten per cent of the energy necessary to cause fracture. It seems, therefore, that crack-spreading by splitting due to atoms occasionally possessing more than the average thermal motion is not likely to be effective at stresses of the order of one-third of those necessary to break the material quickly, as in this case the thermal motion would need to be of the order of ninety per cent of the energy to fracture, that is about nine times the average thermal motion. In addition, a certain amount of co-operation between neighbouring atoms would be necessary. If only one atom at a time acquired sufficient thermal energy to cause splitting, the increase in the loads withstood by neighbouring atoms would not be great, and when the atom had lost its excess thermal motion it would probably return to its previous position of equilibrium and no permanent increase in crack length would result. The probability of sufficient co-operative action of neighbouring atoms is likely to be very small. A quantitative estimate of the rate of crack-spreading by the process described above is probably impossible with our present limited physical knowledge, but it is fairly certain that this process would lead to a much slower decay in strength than the 3 : 1 strength ratio corresponding to a $10^6 : 1$ time ratio found in practice.

§4. DELAYED FRACTURE CAUSED BY APPROACH TO HOMOGENEOUS EQUILIBRIUM

Natural processes tend to a reduction in free energy. A stressed solid can attain a state of lower free energy in a number of different ways, and all are possible factors causing delayed fracture. A possible time effect is connected with the entropy change associated with change in stress. Expressions for the stress coefficient of entropy have already been given. The entropy of a solid can be divided into two parts, a part associated with the vibrations of the atoms and a part associated with their mean positions. Change in the kinetic part may be expected to follow change in stress without any appreciable time lag but, on account of energy barriers, the time for atoms to take up their new equilibrium arrangement may be appreciable. The strength of the material will depend on the atomic arrangement, and so it may change with duration of loading. For unidirectional stress X, the entropy change in an isotropic body is given by

$$\frac{\partial S}{\partial X_T} = V^S \frac{\partial e}{\partial T_X}\left(1 - X \frac{\partial}{\partial X_T}\,(2e_x - \overline{e_y + e_z})\right), \qquad \ldots\ldots(7)$$

where $\partial e/\partial T_X$ is the change in strain with temperature at constant stress and e_x, e_y and e_z are the strain in the directions of the three axes.

Assuming Hooke's law, and ignoring variation in $\partial e/\partial T$ with stress, this expression can be integrated and gives the maximum entropy change with stress as

$$V^S \frac{\partial e}{\partial T}\frac{G}{2},$$

where G is the shear modulus. If this is compared with the entropy change known greatly to reduce the strength of a material (for example the entropy change during melting of crystalline materials with latent heats of evaporation of the same order as mineral glass), it is found to be a small fraction of it, less than 1%. The entropy change given above is the combined kinetic and positional entropy, so that the positional effect has been over-estimated, unless the two parts of the entropy change have opposite signs. It is concluded that the entropy change due to stress in material in homogeneous thermal equilibrium, when stress-free, is not likely to cause an appreciable delayed-fracture effect. It has been assumed that the transition temperatures of any possible phase changes have not been reduced to room temperature by the applied stresses. Phase changes under stress are considered in § 6.

If the material is not in thermal equilibrium, a more appreciable effect may be expected. Solids which have been cooled from the molten state have a higher free energy than they would if they reached the state which is in equilibrium at their actual temperatures, the excess being greater the more rapidly the material is cooled. This is because, due to viscosity, there is not time for all the atoms to reach their stable equilibrium configurations. Mineral glasses, owing to asymmetry of the molecules, have especially high viscosity, and it would be expected that their excess free energy would be relatively high. In such materials there is a tendency for atoms to rearrange themselves, and in a non-crystalline material some atoms are so situated that little energy is required to move them; these can migrate to new positions. As the strength of materials depends on the relative positions and motions of their atoms, their strength, even if stored stress-free, will vary with age. This is found to be the case for freshly made glass rods and fibres. The approach to thermal equilibrium, however, becomes slower as the least stable atoms find positions of greater stability, and after a time the process effectively ceases and strength becomes independent of age. The atomic arrangement is still not that in equilibrium at the temperature, but the energy required to move the least stable atoms from their immediate environment has become large, compared with the average kinetic energy of the atoms, so that migrations are relatively infrequent. When stress is applied, the average work to enable an atom to migrate is reduced, and the kinetic energy of atoms with energy little more than the average is again sufficient to cause migration at an appreciable rate. By increasing the mobility of atoms, stress thus enables the process of approach to equilibrium to continue, and because the strength depends on the atomic arrangement, strength will vary with duration of loading. The increased mobility of atoms under stress may be looked upon as a reduction in internal viscosity.

An indirect effect of the approach to equilibrium is probably very important. Changes in free energy are accompanied by volume changes, and if such volume changes are non-uniform, internal stresses result. A likely cause of the cracks which are the source of weakness of glass is that they are due to tensile stresses set up by non-uniform volume changes in the glass as its atoms rearrange themselves so as to approach equilibrium configuration. The disintegration of a material during crystallization is the extreme form of the same effect. When a material containing initial cracks is stressed, the approach to equilibrium made possible by the stress will be much more rapid in the highly stressed material at the bottom of a crack than elsewhere. The relatively rapid approach to equilibrium in the material at the bottom of the crack will cause differential shrinkage of this material and will increase the stress. It is possible that this process causes an appreciable part of the delayed-fracture effect in glass.

§ 5. DELAYED FRACTURE CAUSED BY APPROACH TO HETEROGENEOUS EQUILIBRIUM

Under this heading two effects will be considered: delayed fracture caused by evaporation of material at the end of the cracks, and the effects of approach to heterogeneous equilibrium with atmospheric components. Evaporation will be considered first.

If data are available for some high temperature, the vapour pressure of a material at room temperature can be calculated from the Clausius-Clapeyron formula and the equation connecting latent heat with temperature. It is also necessary to know ΔC_p (the specific heat change on vaporization) over the temperature range. If the specific volume and coefficient of expansion of the solid are neglected compared with the corresponding quantities for the gas, the vapour pressures at temperatures T_1 and T_2 are related by the formula

$$R \ln \frac{p_2}{p_1} = \int_{T_1}^{T_2} \frac{\left(L_2 - \int_{T_1}^{T_2} \Delta C_p \, dT\right)}{T^2} dT, \qquad \ldots\ldots(8)$$

where L is the latent heat.

Using thermal data at 2000° c. for silica glass and assuming the specific heat of silica gas to be the same as that of CO_2 which has the same number of degrees of primary vibrational freedom as SiO_2 molecules, the vapour pressure at room temperature is estimated from equation (8) to be 10^{-78} dynes/cm^2. The vapour pressure is thus entirely negligible. Due to curvature, the stress-free vapour pressure at the end of a crack will be less than that in equilibrium with a flat surface. If the cracks are elliptical, and if the radius of curvature at the end of the crack is assumed to remain constant as the crack spreads, the effect of curvature on vapour pressure can be shown to be equal to that due to a stress equal to the radial component of the surface tension. If the radius of curvature at the ends of the crack is of the order of molecular dimensions, the vapour pressure of silica glass is further reduced by a factor of the order of 10^{-7}, giving a stress-free vapour pressure at the end of the crack of about 10^{-85} dynes/cm^2. However, on account of the logarithmic relationship between vapour pressure and the square of the stress (see equation (3)), at high stresses such as occur at the ends of cracks, the vapour pressure, and consequently the rate of evaporation, may become appreciable. Equation (3) is based on the equilibrium condition in which evaporation and condensation take place reversibly. Before attempting to calculate the rate of evaporation, the reversibility of evaporation and condensation from a stressed solid needs further examination. Under a hydrostatic stress system, the condensed material must withstand the full pressure, and conditions are thus truly reversible, but in the case of a surface having a tension stress in its plane, material condensing on the surface could deposit so as to be stress-free. In the case of a crystalline material strained to a small extent, it is possible that the deposited material would be acted on by the full stress. On account of the difference in interatomic spacing between stressed and unstressed material, if the deposited material is unstressed its atomic arrangement must be somewhat irregular, and this irregularity may entail a higher free energy than if the deposited material withstood the full stress. As the stress is increased, however, the free energy of an amorphous atomic arrangement will become less than that of the stressed crystalline arrangement. It would thus be expected that an unstressed non-crystalline layer would cover the surface of the material. In an amorphous material it would be expected that the deposited layer would be unstressed whatever stress was in the parent material. If, therefore, a stressed material is allowed to remain in contact with its saturated vapour, its surface may be expected to become contaminated with unstressed material and consequently the vapour pressure will fall. If the radius of curvature at the end of the crack in a stressed solid is of the order of molecular dimensions, the highly stressed area will be small compared with the mean free path of the molecules of the gas (except at very high vapour pressures), and as the highly stressed area is surrounded by comparatively unstressed material, with negligible vapour pressure, all atoms escaping from it will condense on the surrounding surface and the material at the end of the crack may be expected to be uncontaminated. The vapour pressure calculated on the basis of a fully stressed surface may therefore be expected to give the correct rate of evaporation from a surface at the end of a crack. If $2a$ is the length of the crack, the rate of spread of the crack due to evaporation estimated from kinetic theory is

$$\frac{da}{dt}=\frac{V^S}{V^G}\frac{\bar{c}}{4}, \qquad \ldots\ldots(9)$$

where V^S and V^G are the specific volumes of solid and gas and $\bar{c}$ is the root mean square of the velocity of the gas molecules, which for a perfect gas at a given temperature can be estimated from kinetic theory. V^G depends on the vapour pressure, which in turn depends on the stress at the root of the crack. This stress for a given average stress can be expressed in terms of the shape of the crack (in principle at least) if the shape of the crack is known. Failure will occur when the crack attains a length such that the stress at its root is the maximum the material can withstand. In practice, failure would be virtually immediate if the vapour pressure became high. Assuming the cracks to be elliptical, the stress at their ends can be expressed in terms of the average stresses, using Inglis's (1913) solution. As the ratio of the major to minor axis of the elliptic crack is large, the stress is sufficiently accurately given by

$$X=2f\sqrt{\frac{a}{r}}, \qquad \ldots\ldots(10)$$

where X is the maximum stress at the end of the crack, f is the average stress to which the material is subjected, $2a$ is the crack length, and r the radius of curvature at the end of the crack. Using equation (3) to give vapour pressure in terms of stress, and assuming the vapour to be a perfect gas, equation (9) becomes.

$$\frac{da}{dt}=\frac{V^S}{\sqrt{2\pi RTM}}\,p_0 e^{\frac{2f^2V^Sa}{ERTr}}. \qquad \ldots\ldots(11)$$

Integration gives

$$\frac{1}{p}-\frac{1}{p_i}=-\frac{2f^2(V^S)^2}{ERTr\sqrt{2\pi RTM}}\,t, \qquad \ldots\ldots(12)$$

where p_i is the vapour pressure at the end of the crack when the average stress f has just been applied. Failure would occur when, due to crack-spreading, the stress at the end of the crack reaches the maximum the material can withstand. If this is estimated by the methods outlined in § 2, the latent heat of evaporation does not become zero until the vapour pressure is high. At failure, therefore, $1/p$ in equation (12) may be neglected compared with $1/p_i$, and this equation thus leads to

$$\left(\frac{A}{f_0^2t_0}\right)^{\frac{1}{f_0^2}}=\left(\frac{A}{f^2t}\right)^{\frac{1}{f^2}}, \qquad \ldots\ldots(13)$$

where f and t are the average stress and time to fracture in terms of f_0 and t_0, which are corresponding quantities in an arbitrary basic state. The constant A is given by

$$A=\frac{ERTr\sqrt{2\pi RTM}}{2(V^S)^2p_0}. \qquad \ldots\ldots(14)$$

Assuming r to be the cube root of the volume occupied by a molecule, equation (13) has been applied to silica glass (molecule $=(SiO_2)_1$) having an average strength of about 13,000 lb./sq. in. for a ten-second duration of loading.

If the delayed fracture of this material is wholly attributed to spread of cracks by evaporation and the given values of f_0 and t_0 are put in equation (13), the times to fracture at other stresses can be computed. For example, the time to fracture at a stress 5% less than that giving fracture in 10 sec. is calculated to be of the order of sixty years, whereas the corresponding experimental time is about 30 sec. If only a part of the delayed-fracture effect is attributed to evaporation, the corresponding calculated fracture time will be increased. It seems reasonable, therefore, even having regard to the assumptions and approximations introduced into the calculations, to conclude that the time effect for mineral glass stressed in tension, attributable to evaporation, is very small.

A much more important cause of delayed fracture is atmospheric attack of the glass surface. The decay of ancient stained glass is evidence that this takes place even at room temperature. At higher temperatures glass is readily soluble in water. Barus (1891), when measuring the compressibility of water at elevated temperature, found that the inside of his capillary tube (made of lead glass) rapidly dissolved in water at 185° c. At 350° c. even pure silica glass is readily soluble. The present author, in collaboration with Pearson, has recently investigated the delayed bending fracture of round soda-glass rods of $\frac{1}{4}$ in. diam. when sealed in flexible metal evacuated tubes. The slope of the curve of (stress v. log time to fracture) obtained for the material tested *in vacuo* was about half that for material tested in air at 75% R.H.*

Preston (1942) and Orowan (1944) had previously attributed delayed fracture to atmospheric attack, the former from general and the latter from theoretical considerations. Preston did not give any detailed explanation of the process. Orowan's theory has already been discussed. Instead of attributing delayed fracture to the reduction of surface tension caused by atmospheric attack, it seems better to attribute it directly to loss in strength of the material at the root of the crack. Due to the concentration of strain energy, this material has a much higher free energy than normal unstressed glass and is, therefore, much more chemically active. Continuous reaction with atmospheric constituents is possible even without solution, if the stress at the base of the crack is greater than the strength of the glass-atmospheric constituents complex formed there; for in this case the crack will extend continuously and expose uncontaminated glass to the atmosphere. It is however convenient, and possibly not inaccurate, to regard the process as one of solution of the glass in atmospheric constituents, because in this case the expression already obtained for the delayed fracture caused by evaporation may, with altered constants, be applicable, as solubility in dilute solutions is known to be proportional to vapour pressure. The theory involved in deducing equation (13) gives, for the length of crack (a) in material subjected to constant load, an expression of the form

$$a = -B \log(C - Dt). \qquad \qquad \ldots\ldots(15)$$

For crack-spreading by evaporation the constants C and D are very small numbers (of the order of 10^{-40}). For example, if the time to break was 10^5 sec., C might be 10^{-40} and D 10^{-45}. After 10^5 sec., the crack length would be infinite, but after 9×10^4 sec. the crack length would have increased by only $2\frac{1}{2}$%. The crack length is thus effectively constant until a critical time is reached at which it increases extremely rapidly. Crack-spreading by solution would be expected to be much more rapid than crack-spreading by evaporation, and if an equation similar to expression (15) holds, the constants C and D would be much greater, but the crack length may still remain sensibly constant until just before fracture, so that crack-spreading in mineral glass by solution in atmospheric constituents may be

---

* In recently published work (1946) Baker and Preston found in some cases that when tests were made *in vacuo* no delayed fracture occurred during the time range 10^{-3} to 10 sec.

consistent with the results of Murgatroyd and Sykes quoted in § 1. It is of interest to mention here that rock salt has been found to be appreciably stronger when broken in water (Joffé, 1935). A likely explanation is that, on account of the high solubility, even when unstressed, the highly stressed material at the ends of the cracks becomes covered with a layer of unstressed material deposited from solution. If this happens, on account of the high curvature at the ends of the cracks, the vapour pressure, and hence solubility, will be less than elsewhere, and there will be a tendency for material to deposit at the ends of the cracks and thus to heal them.

§ 6. THE EFFECT OF POSSIBLE PHASE CHANGES UNDER STRESS

At the beginning of § 4, effects due to the entropy change associated with change of stress were discussed, and in materials in thermal equilibrium when stress-free the effects were considered to be unimportant. Much more important entropy changes may occur if, due to stress, the transition temperatures of phase changes is reduced to that of the test temperature, so that phase changes tend to occur. The two phases will be referred to as the high-temperature phase (H.T.P.) and the low-temperature phase (L.T.P.). The sign of the latent heat of transition is taken from the sign of the entropy difference between the H.T.P. and the L.T.P.

The stress coefficient of transition temperature can be calculated by the process used to calculate the stress coefficient of vapour pressure. The result depends on the stress system withstood by the two phases. If the high-temperature phase is a mobile liquid, it can withstand no stress other than hydrostatic pressure. One case to be considered, therefore, is that of a stress-free H.T.P. If both phases can withstand stress, theoretical conditions can be devised so that the relative stresses in the two phases have any given values, but for simplicity the other case treated here is that in which the two phases have the same stresses.

If the H.T.P. is unstressed and the transition takes place on the *mm* face of a cube, and if $kl \neq mm$, the stress coefficient of transition temperature is

$$\frac{\partial T}{\partial X_{kl}} = \frac{\Sigma X_{ij} \dfrac{\partial e_{ij}}{\partial X_{kl}} - X_{mm} \Sigma_{ii} \dfrac{\partial e_{ii}}{\partial X_{kl}}}{-\dfrac{L}{TV^S} - \Sigma X_{ij} \dfrac{\partial e_{ij}}{\partial T} + \dfrac{X_{mm}}{V^S} \dfrac{\partial V^S}{\partial T}}. \qquad \ldots\ldots(16)$$

If $kl = mm$ we have

$$\frac{\partial T}{\partial X_{mm}} = \frac{-1 + \Sigma X_{ij} \dfrac{\partial e_{ij}}{\partial X_{mm}} - X_{mm} \Sigma_{ii} \dfrac{\partial e_{ii}}{\partial X_{mm}}}{-\dfrac{L}{TV^S} - \Sigma X_{ij} \dfrac{\partial e_{ij}}{\partial T} + \dfrac{X_{mm}}{V^S} \dfrac{\partial V^S}{\partial T}}. \qquad \ldots\ldots(17)$$

If the H.T.P. carries the same stress as the L.T.P. we have for $kl \neq mm$

$$\frac{\partial T}{\partial X_{kl}} = \frac{\Delta \left[V \Sigma X_{ij} \dfrac{\partial e_{ij}}{\partial X_{kl}} - X_{mm} \dfrac{\partial V}{\partial X_{kl}} \right]}{\dfrac{L}{T} \pm \Delta \left[X_{mm} \dfrac{\partial V}{\partial T} - V \Sigma X_{ij} \dfrac{\partial e_{ij}}{\partial T} \right]}, \qquad \ldots\ldots(18)$$

and for $kl = mm$

$$\frac{\partial T}{\partial X_{mm}} = \frac{\Delta\left[V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial X_{mm}} - X_{mm}\frac{\partial V}{\partial X_{mm}} - V\right]}{\frac{L}{T} + \Delta\left[X_{mm}\frac{\partial V}{\partial T} - V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial T}\right]}, \qquad \ldots\ldots(19)$$

where the symbol Δ indicates the difference between the corresponding quantities in the H.T.P. and L.T.P.

For unidirectional stress in an isotropic material, the cases for the stress-free H.T.P. reduce to

$$\frac{\partial T}{\partial X} = \frac{X/E}{-\frac{L}{TV^s} - \frac{X\partial e}{\partial T}}, \qquad \ldots\ldots(16a)$$

$$\frac{\partial T}{\partial X} = \frac{1 - 2\nu X/E}{-2X\frac{\partial e}{\partial T} + \frac{L}{TV^s}}. \qquad \ldots\ldots(17a)$$

For the stressed H.T.P. case, the equations for unidirectional stress are

$$\frac{\partial T}{\partial X} = \frac{\Delta\left[V\frac{X}{E}\right]}{\frac{L}{T} - \Delta\left[VX\frac{\partial e}{\partial T}\right]}, \qquad \ldots\ldots(18a)$$

$$\frac{\partial T}{\partial X} = \frac{\Delta\left[2X\nu\frac{V}{E} - V\right]}{\frac{L}{T} + 2X\Delta\left[V\frac{\partial e}{\partial T}\right]}. \qquad \ldots\ldots(19a)$$

In these equations, the latent heat, Young's modulus, specific volume and coefficient of expansion are functions of stress and temperature. The change in latent heat of melting with stress can be derived by methods similar to those used to compute the stress coefficient of the latent heat of evaporation.

It can be seen that at low stresses the stress coefficient of transition temperature is much greater for transition off the face normal to the particular stress than for transition off other faces. The case represented by equations (17) and (17 *a*), however, is of no interest in connection with tensile stresses, as the stress must continue to be applied to the surface of the solid as this recedes, due to melting, and the liquid must remain stress-free; but it has been suggested by Goranson (1940) that it is important in connection with the flow of materials in compression, flow being due to melting under stress. For aluminium a compression stress of the order of 65 tons/sq. in. would reduce the melting temperature to room temperature. Considering the other cases, the stress-free H.T.P. forming off an unstressed face (equation (16 *a*)) leads to a reduction in transition temperature with tension or compressive stress. The stressed H.T.P., forming off a stress-free face (equation (18 *a*)) leads to an increase in transition temperature for tension or compressive stress as $\Delta[V/E]$ is positive for most materials. The stressed H.T.P. forming off the stressed face (equation (19 *a*)) leads to a reduction in transition temperature for tension stress and an increase for compression stress. An idea of the magnitude of the reduction in transition temperature given by equation (19 *a*) may be obtained by computing the temperature coefficient of stress at stresses small compared with E. At low stress the equation reduces to

$$\frac{\partial T}{\partial X} = -\frac{T\Delta V}{L}. \qquad \ldots\ldots(19b)$$

If this formula is applied to a crystalline material, with about the same stress-free "melting" temperature and the same Young's modulus as soda glass, for example aluminium, the estimated stress at which the melting temperature equals room temperature is about 350 tons/sq. in.

It is seen, therefore, that for solids initially in thermal equilibrium, high stresses such as occur at points of stress concentration may be sufficient to cause changes in the stable phase. The phase changes will in many cases take place gradually, and could therefore cause delayed fracture.

Mineral glass at room temperature is not in thermal equilibrium, the high-temperature phase persisting at temperatures below the transition point. The effect of stress on change in the transition point may not therefore be important. If stress lowers the transition temperature, its direct effect would be to reduce the tendency for the glass to crystallize, although the indirect effect (dealt with in § 4) of reducing viscosity and so assisting the glass to crystallize may be the more important effect. If stress raises the transition temperature, the free energy reduction on crystallizing will be increased, and for this reason the rate of crystallization will increase, but this would not be expected to be an important direct effect.

The effect of change in transition temperature caused by stress is not, therefore, expected to be important for glass, although it may be important for metals and other materials whose liquid phases have relatively low viscosity. An unknown factor in this connection is the rate of phase transition. In some cases it may be so slow that the transitions are indefinitely delayed.

§ 7. CONCLUSIONS

1. Previous attempts to explain delayed fracture in glass are not consistent with all known facts.

2. In materials having atomic constitution, Griffith's criterion does not give the stress at which immediate fracture occurs. It gives the lowest stress at which a crack in a material in homogeneous thermal equilibrium may start to spread by a process of splitting. The rate of crack-spreading depends on the rate at which thermal motions overcome energy barriers, but this is considered to be too slow to cause an appreciable delayed-fracture effect.

3. Immediate fracture does not occur until the stress at the end of a crack equals the maximum the material can withstand. The maximum stress is that at which the average thermal motion just overcomes the average energy barrier. The nearest estimate of this stress obtainable from thermal data is that it is the intensity of the applied stress system at which the latent heat of evaporation is zero.

4. In material in homogeneous equilibrium when stress-free, entropy changes due to stress are not likely to cause appreciable delayed-fracture effects.

5. Glass is not in homogeneous thermal equilibrium. Immediately after manufacture, glass tends to approach equilibrium, but the process soon effectively ceases because of the high internal viscosity. Stress reduces the internal viscosity and thus enables the approach to equilibrium to continue. Changes will take place much faster in the highly stressed material at the root of the crack than elsewhere, and as approach to equilibrium results in volume shrinkage, the tensile stress at the ends of cracks increases and causes them to spread. This process is expected to result in an appreciable delayed-fracture effect.

6. Crack-spreading by evaporation of the material at ends of cracks is too slow to give an important delayed fracture effect.

7. Atmospheric attack causes an important delayed-fracture effect. Instead of regarding the effect as due to reduction in surface tension, it is better to attribute it to weakening of the material at the end of the crack. The free energy of this material is much greater than that of normal glass on account of the strain energy concentration, and thus the material is more chemically active. A complex of glass and atmospheric constituents will be formed at the end of the crack. The crack will extend continually if the strength of this complex during or after its formation is less than the load imposed on it.

8. Change, due to stress, in the stable phase at room temperature is not likely to be important for glass, as in this material the high-temperature phase persists at temperatures below the transition point. Phase changes under stress may have important effects on the behaviour of other materials.

LIST OF SYMBOLS

X	Unidirectional stress.
X_{ij}, X_{kl}	Generalized stress.
e	Unidirectional strain.
e_{ij}, e_{kl}	Generalized strain.
p	Pressure.
V	Volume.
V^G, V^L, V^S	Specific volumes of gas, liquid and solid.
E	Young's Modulus.
G	Shear modulus. Also superscript for gas.
S	Entropy. Also superscript for solid.
T	Absolute temperature.
L	Latent heat of isothermal transition.
R	Gas constant.
C_p	Specific heat at constant pressure.
Δ	Difference.
a	Semi-crack length.
$\bar{c}$	Root mean square of velocity of gas molecules.
r	Radius at rooted crack.
f	Average stress.
M	Molecular weight.

REFERENCES

ANDERBGG, F. O., 1939. *Ind. Eng. Chem.*, **31**, 290.
BARUS, C., 1891. *Amer. J. Sci.*, **41**, 110.
GORANSON, R. W., 1940. *J. Chem. Phys.*, **8**, 323.
GRENET, L., 1899. *Bull. Soc. Encouragement*, **4**, 839.
GRIFFITH, A. A., 1920. *Phil. Trans.*, **221**, 163.
GURNEY, C. and PEARSON, S., 1945. Awaiting publication.
HOLLAND, A. J. and TURNER, W. E. S., 1940. *J. Soc. Glass Tech.*, **24**, 46.
HOUWINK, R., 1937. *Elasticity Plasticity and Structure of Matter* (Cambridge: The University Press).

Inglis, C. E., 1913. *Proc. Inst. Naval Architects*, **45**, 219.
Joffé, A., 1935. *Rep. Int. Conf. Physics* (London : Phys. Soc.).
Murgatroyd, J. B., 1944. *J. Soc. Glass Tech.* **28**, 406.
Murgatroyd, J. B. and Sykes, F. S., 1945. *Nature, Lond.*, **156**, 716.
Orowan, E., 1944. *Nature, Lond.*, **154**, 341.
Preston, F. W., 1942. *J. Appl. Phys.*, **13**, 623.
Preston, F. W., Baker, T. C. and Glathart, J. L., 1946. *J. Appl. Phys.*, **17**, 162.

CAVITY RESONATORS FOR MEASUREMENTS WITH CENTIMETRE ELECTROMAGNETIC WAVES

By B. BLEANEY, J. H. N. LOUBSER and R. P. PENROSE,
Clarendon Laboratory, Oxford

MS. received 12 *June* 1946 ; *in revised form* 10 *October* 1946

ABSTRACT. A wave-meter for wave-lengths of about a centimetre, with an accuracy of 1 to 2 parts in 10 000, is described, based on a new system of coupling to the H_0 mode in resonant cavities which avoids the excitation of other modes. An indirect effect due to simultaneous resonance in two different modes is discussed in relation to the measurement of absorption by resonant cavities.

The dielectric constants of six non-polar liquids have been determined, by means of wave-meters of this type, at wave-lengths of 3·2 and 1·35 cm. The values obtained at the two wave-lengths agree in all cases within 3 parts in 10 000, which is consistent with the estimated experimental error, and do not differ appreciably from the accepted low-frequency values.

The temperature coefficient of the dielectric constant is compared with that calculated from the dilatation of the fluid. The power factors vary between 10^{-4} and $2 \cdot 10^{-3}$; the higher values may be due to traces of polar impurity.

PREFACE

The resonant circuit has always played an essential part in apparatus for radio-frequency measurements, and at centimetre wave-lengths its rôle is no less predominant. Its form is, however, very different from the lumped circuit where the electric and magnetic fields are isolated in the capacity and inductance. A hollow resonator is used from which the loss of energy by radiation is negligible, and the dissipation in resistive loss is very small. The ratio of stored energy to energy lost per cycle is therefore large, and the magnification factor Q is high, of the order of 10^4. The resonance is correspondingly sharp, and the wave-length in the cavity can be determined with great accuracy. The resonant cavity is thus eminently suitable for use as a wave-meter, and for the determination of the properties of low-loss dielectrics at centimetre wave-lengths.

The difference between the lumped circuit and the resonant cavity is not only in form but also in nature. The concepts of current and voltage are replaced by those of magnetic and electric field, and the concept of impedance is therefore only of subsidiary importance. In Part I of this paper the theory of a new method of exciting one particular mode of resonance, the H_0, in the cavity, is developed by considering the fields in the cavity and in the wave-guide feeder. A precision wave-meter based on these principles is described, which covers the wave-length band from 1·15 cm. to 1·55 cm.

In Part II the application of such wave-meters to the measurement of the dielectric properties of six non-polar liquids at wave-lengths of 3·2 cm. and 1·35 cm. is described, and the results are briefly discussed.

PART I.—DESIGN OF WAVE-METER

§ 1. INTRODUCTION

It is convenient to divide hollow resonators into two types: coaxial line resonators and wave-guide (cavity) resonators. In the coaxial line the waves of the normal type travel with the same velocity as in free space, and the wave-length may be directly determined from the distance between successive points of resonance, which will occur at intervals of half a wave-length. To ensure that only waves of the normal type are present, the diameter of the outer conductor of the coaxial line must be small compared with the wave-length. At the shorter wave-lengths this restriction results in considerable mechanical difficulties in the design of coaxial line resonators, and also increases the ratio of the dissipation of energy per cycle to the stored energy. The Q value is therefore smaller and the accuracy of setting is reduced. These difficulties are overcome by the use of wave-guide resonators, with which the problem becomes the elimination of all but the desired wave-type. One simple solution is the use of a wave-guide in which only one mode of propagation, the lowest, or H_1 mode, is possible. The diameter * of the wave-guide must, for this purpose, lie between 0·59 and 0·77 of the wave-length. The useful wave-length range is therefore less than 20%. Moreover, a much more serious difficulty arises at wave-lengths of a centimetre or less. In a wave-guide the velocity of propagation depends on the ratio of the wave-length to the diameter of the guide. The latter must therefore be constant and known with high precision; when the diameter is only a few millimetres the machining tolerances demanded become prohibitive. These difficulties may be overcome by the use of a more suitable type of wave than the H_1.

Although an unlimited number of modes of propagation are available in a wave-guide of sufficient size, the choice of mode for an accurate wave-meter is readily narrowed down to one, the H_0. The especial property of this mode—of giving a low dissipation of energy on the walls of the guide—has long been recognized. In addition it has the following advantages:—

1. There is no flow of current in the radial direction on the end walls. The cavity may therefore be tuned by a non-contact piston without leakage of energy past the piston.

2. The H_0 mode is a singlet mode, whereas all other modes (except E_0) are doublets, corresponding to the possibility of either cosine or sine distribution of field with respect to azimuth. The degeneracy of these doublet modes is removed by a slight ellipticity, and double resonances may appear. The H_0 mode is stable against small irregularities in the guide.

---

* The advantages of circular guide (e.g. in machining) are so obvious that discussion of other shapes is omitted.

3. The diameter of the guide must be greater than 1·22 wave-lengths, and with the coupling system described below, a diameter as large as $2\frac{1}{2}$ wave-lengths has been used. For the shorter wave-lengths this makes the machining tolerances much easier.

A large size of guide permits, of course, the propagation of a number of other wave-types, which may give rise to undesired and misleading resonances. To remove the ambiguities thus caused it has hitherto been customary to attempt to damp-out the undesired resonances by placing " lossy " material (e.g. resistive wires) at suitable positions within the resonator. Such methods are only partially effective, and have the disadvantage of introducing irregularities in the cavity.

A more fundamental approach to the problem suggests the elimination of undesired resonances by the use of a properly designed system of coupling the resonator to the feeder. A coupling unit such as a probe or loop, or, preferably, a hole, can only feed energy into modes of resonance which have the appropriate field components at the coupling unit. Modes in which these components are zero will not be excited. This principle is combined, in the coupling system described below, with the use of a double symmetrical feed which eliminates all modes whose fields vary with $\genfrac{}{}{0pt}{}{\cos}{\sin} n\phi$ round the periphery if n is an odd integer. This elimination depends, of course, on an exact balance between the two halves of the feed, which is obtained by the use of two symmetrical holes of equal diameter. It is for this purpose that holes are greatly to be preferred to probes or loops, since the balance is automatically achieved, whereas probes or loops would require delicate adjustment by hand which would probably be frequency-sensitive.

The correct location of the coupling holes in the feeder is also important. When the feeder is a rectangular H_1 wave-guide, as is usual at the shorter wave-lengths, the obvious position for the coupling holes is on the narrow side of the guide, since the only field component here is a longitudinal magnetic field. If the coupling holes are small, this field will not be seriously disturbed by their presence, and only modes in the resonator which have a parallel magnetic-field component will be excited. On the broad side of the guide there are three field components, and the number of modes likely to be excited in the resonator is correspondingly greater. Thus as a general rule coupling out of this side is to be avoided.

The theory of this system of coupling is developed in the following sections, and two H_0 wave-meters based on these principles are described. Freedom from undesired resonances over a considerable wave-length range is obtained, the performance in this respect fully confirming the predictions.

§2. THEORY OF THE COUPLING SYSTEM

The field components at the end wall of a cylindrical cavity are (cf. Lamont, *Wave-guides* (Methuen), p. 76):

for H-waves	for E-waves
$H_z=0$	$H_z=0$
$H_\rho=(2\pi k/\lambda_g)J_n'(k\rho)\cos n\phi$	$H_\rho=-j(2\pi n/\lambda_a\rho)\,J_n(k\rho)\sin n\phi$
$H_\phi=-(2\pi n/\lambda_g\rho)J_n(k\rho)\sin n\phi$	$H_\phi=-j(2\pi k/\lambda_a)\,J_n'(k\rho)\cos n\phi$

for H-waves (*cont.*)	for E-waves (*cont.*)
$E_z=0$	$E_z=k^2J_n(k\rho)\cos n\phi$
$E_\rho=0$	$E_\rho=0$
$E_\phi=0$	$E_\phi=0$

where the nomenclature is the same as that used by Lamont, viz. (z, ρ, ϕ) are cylindrical polar co-ordinates, k is defined by $J_n'(ka)=0$ where a is the radius of the resonator, and J_n is the Bessel function of order n. λ_g, λ_a are the wave-lengths in the cavity and in free space respectively.

The only field existing along the narrow side of an H_1 rectangular wave-guide is a longitudinal magnetic field. If, therefore, coupling holes in the side wall open into the end of a cylindrical cavity along an axial plane, only the radial component of the magnetic field in the cavity will be coupled to the field in the wave-guide. Thus we need only consider the behaviour of H_ρ in the equations above. Since H_ρ varies as $\cos n\phi$, the radial fields at diametrically opposite points will be of the same sign if n is even, and of opposite sign if n is odd. This is illustrated by the sketches (figure 1) of the magnetic fields H_0, H_1 and H_2. It is possible, therefore, to avoid the excitation of any E-wave or H-wave of odd order by using two equal coupling holes spaced half a wave-length apart in the rectangular wave-guide, and placed at diametrically opposite points in the cylindrical cavity.

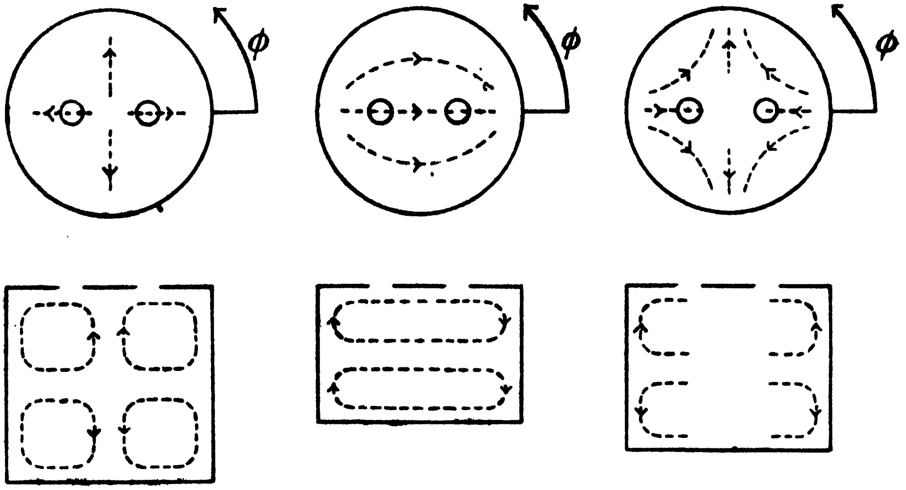
Table 1. Ratio of diameter to cut-off wave-length for the principal modes in a cylindrical resonator

H-waves	H_{n_1}	H_{n_2}	H_{n_3}
H_0	1·2197	2·2331	
H_1	0·5861	1·6970	2·7172
H_2	0·9722	2·1346	3·1734
H_3	1·3373	2·5513	
H_4	1·6926	2·9547	
H_5	2·0421		
H_6	2·3877		
H_7	2·7304		
H_8	3·0709		

E-waves	E_{n_1}	E_{n_2}	E_{n_3}
E_0	0·7655	1·7571	2·7546
E_1	1·2197	2·2331	
E_2	1·6347	2·6793	
E_3	2·0309	3·1070	
E_4	2·4154		
E_5	2·7920		
E_6	3·1628		

***Accuracy.* ±1 in last place (or better). All modes for which *d/c* is less than 3 are included.**

Table 1 gives values of d/λ_0 for various wave types. In a typical H_0 cavity, d/λ_0 is 1·52 at the centre of the band; waves for which d/λ_0 exceeds this value are therefore beyond cut-off. Of the permissible modes of resonance in this size of resonator, H_1, H_3 and E_1 are not excited for reasons already discussed; neither is E_0, for it has no radial component of H. The only undesired modes which are liable to occur are therefore H_2, H_4 and E_2; of these, the last two do not arise except at the shorter wave-lengths. The coupling to the higher-order waves is in any case weak, since their radial components of magnetic field are small at the coupling holes. In a transmission-type wave-meter it is possible to eliminate the H_2 resonances by suitable placing of the output coupling, since the field components vary with $\sin 2\phi$ or $\cos 2\phi$. It is convenient to use an output coupling of the usual type—a single hole in the side of the wave-meter barrel a quarter wave-length from the end wall—feeding a rectangular wave-guide in which the plane containing the magnetic field passes through the axis of the cylindrical cavity. The output guide then couples to the H_z field in the cavity; for H_2, this field is zero at an angle of 45° to the input guide. Thus if the output is oriented at this angle (as in figure 2) it will not couple to any E-wave, since by definition these have no longitudinal component of magnetic field.



H_0 H_1 H_2

Figure 1. Magnetic fields in cylindrical cavity.

Band-width of the coupling system

Since the input coupling makes use of two holes spaced half a wave-length apart, its performance will not be so good at wave-lengths other than that for which it was designed. The diminution in its efficiency is determined solely by the change of wave-length in the input wave-guide; change of wave-length in the resonant cavity has no effect since the two coupling holes are placed in the end wall. It is easily seen that if l is the distance between the coupling holes, the coupling to waves of order $2n$ in the cavity is proportional to $\sin \pi l/\lambda_g$ and that to waves of order $2n+1$ is proportional to $\cos \pi l/\lambda_g$. Thus when $l=\lambda_g/2$, the coupling to the wanted waves is a maximum while that to the unwanted waves is zero. To find the useful band-width of the device, we take as a somewhat arbitrary criterion that the coupling to the $2n$ modes (as measured by a square-law detector) shall be at least ten times that to the $2n+1$ modes, assuming that the coupling is determined mainly by the sine and cosine terms. We have, then, at the end of the band

$$\tan^2 \pi l/\lambda_g = 10,$$

the solution of which gives approximately $\lambda_g = 2{\cdot}5l$ or $1{\cdot}6l$. Taking the typical value 1·25 for the ratio λ_g/λ at the centre of the wave-band λ_0, one finds that the instrument should work satisfactorily at wave-lengths between $1{\cdot}15\lambda_0$ and $0{\cdot}90\lambda_0$—a 25% band.

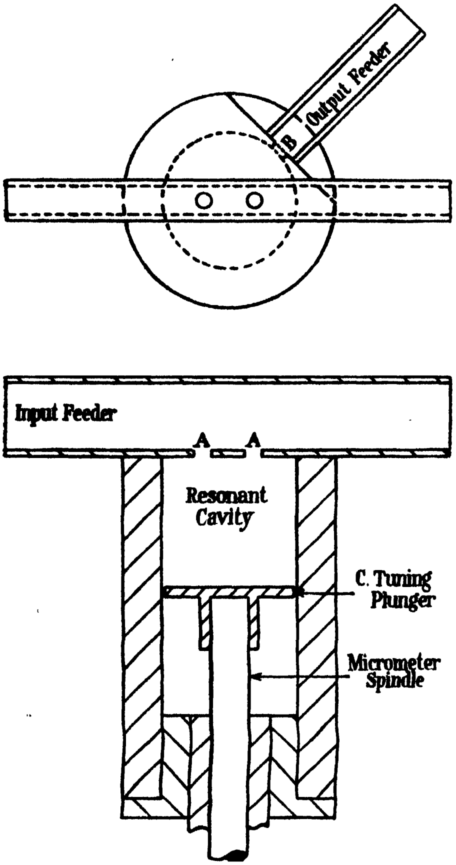


Figure 2. H_0 resonator.

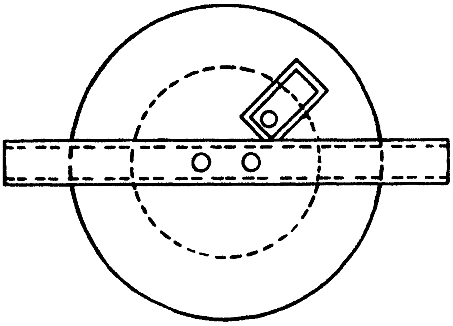


Figure 3. Alternative arrangement with output coupling on end.

§ 3. EXPERIMENTAL CONFIRMATION

The conclusions drawn in the previous paragraph were tested in the region of 1 cm. wave-length. A cylindrical cavity was made with an internal diameter of 19 mm. (figure 2); it was coupled to the H_1 rectangular wave-guide by two holes (A A) spaced 7·7 mm. apart, which in this particular guide corresponds to a half wave-length for a frequency of 24 000 Mc./sec. (1¼ cm. wave-length). The output coupling hole B was drilled in the circular barrel at a point one quarter of a wave-length (in the cavity) from the end, and oriented at 45° to the axis of the input coupling. The cavity was tuned by a plunger C driven by a

$2\frac{1}{2}''$ micrometer head with 25 mm. travel, manufactured by Messrs. Moore and Wright. The coupling holes were adjusted in size by experiment until adequate power was available in the detector, a crystal.

In order to make a direct comparison of the performance of the new coupling system with that of the conventional coupling system, a fourth coupling hole was drilled similar and diametrically opposite to the hole B. This fourth hole and B form a coupling system of the usual type, and the resonator could be quickly transferred between this system and the new coupling system so that they could be compared under identical conditions. This comparison was made at $1\frac{1}{4}$ cm. wave-length. With the conventional coupling system, several modes of resonance were easily detected, one of them, H_3, being over twice as strong as the desired H_0 resonance! On the other hand, with the new coupling system only one undesired mode could be detected by a systematic search. This was the H_2 mode, and it gave a response smaller than that of the H_0 mode by a factor of several hundred. Its presence was probably due to the fact that in the first model the output coupling hole was not quite at 45° to the input holes; in any case, such a small response would pass unnoticed in ordinary use.

The resonator was next tested at various points in a range of wave-length 1·08 to 1·48 cm. No difficulty was experienced with undesired modes; the performance of the coupling system was satisfactory over the 25% band predicted by the considerations of § 2.

§ 4. THE WAVE-METER

During work on absorption in gases near 1 cm. wave-length, the need arose for a wave-meter with an accuracy of one or two parts in ten thousand. In the first experiments, a resonant cavity similar to that described in the previous section was used to measure wave-length, but its accuracy was only about one part in a thousand. The limitation of accuracy was almost entirely due to the size of the barrel. The effect of errors in the diameter of the barrel may be estimated from the equation

$$\frac{d\lambda_c}{\lambda_c}=\frac{d\lambda}{\lambda}\cdot\frac{\lambda_c^2}{\lambda^2} \qquad \ldots\ldots(4.1)$$

obtained by differentiation of the usual wave-guide equation. Here λ_c is the critical wave-length of the cavity (= 0·82 diameters) and λ is the free-space wave-length of the radiation. If $d\lambda/\lambda$ is taken as 10^{-4}, the diameter of the barrel as 19 mm., and $\lambda = 1\frac{1}{4}$ cm., the allowable error in the diameter is only 3 microns. With the resources available in this laboratory the diameter could neither be machined nor measured to this accuracy.

The solution of this difficulty is suggested by the form of equation (4.1), in which the allowable error in the diameter depends on the cube of λ_c. If the diameter of the barrel is increased to 30 mm., the tolerance becomes 12 microns, which is within the limits of good instrument making. The disadvantage of using so large a diameter is, of course, that it permits the propagation of many modes (16 at $1\frac{1}{4}$ cm. wave-length). The new coupling system eliminates, by its symmetry and the use of the 45° angle between input and output, all but the modes for which n is a multiple of 4. The fields of the modes with high values of n are concentrated towards the periphery and are relatively weak near the axis, corresponding to the fact that in the expansion of $J_n(x)$ the lowest term is of the order x^n. Thus in a large diameter cavity, the coupling to modes such as E_4 and H_4 is small because the distance of the input coupling holes from the axis of the cavity is only a small fraction of the radius of the cavity.

The drilling of the output coupling hole in the barrel, and the machining of a flat surface on the outside for the attachment of a wave-guide, are liable to cause an undesirable distortion of the barrel. This may conveniently be avoided by placing the output coupling on the end, again at 45° to the input. Such an arrangement is shown in figure 3. The output hole is located at a distance from the axis equal to 0·55 times the radius of the cavity, where the radial overtone (H_{02}) of the desired mode has zero radial magnetic field. This eliminates the overtone, which otherwise would give a resonance of about the same intensity as the fundamental.

A wave-meter fed in this manner, and with a cavity diameter of 30 mm., has been constructed from copper, and has been very satisfactory. It has been used for wave-lengths ranging from 1·15 cm. to 1·55 cm. without any ambiguities from other modes. The plunger is driven by a 2½-inch micrometer head, graduated to 0·002 mm., which is normally read to half a division. The consistency of the measurements obtained with the instrument is illustrated by the two following sets of readings for the position of resonance, where a fourth decimal place has been determined by estimation of fractions of a small division:—

Micrometer reading	Half wave-length	Micrometer reading	Half wave-length
24·1984		23·8437	
	7·3459		7·2573
16·8525		16·5864	
	7·3449		7·2560
9·5076		9·3304	
	7·3440		7·2558
2·1636		2·0746	

The slight increase in the half wave-length at the higher micrometer readings is consistent with a slight taper in the barrel (less than 0·01 mm. in the diameter) which was just observable by direct measurement. If, in general, an error of 0·002 mm. in reading the wave-length in the cavity is allowed, the corresponding error in the free-space wave-length is about 1 part in 10000. From the width of the resonances the Q of the cavity was estimated as 20000; the positions of resonance could therefore easily be located to within 0·001 mm. The error in the measurement of the diameter of the barrel should not exceed 0·01 mm. An overall accuracy of 2 parts in 10000 in the wave-length may therefore be expected.

§ 5. INDIRECT EFFECT OF OTHER MODES OF RESONANCE

Examination of the performance of the wave-meter showed that at certain wave-lengths one of the H_0 resonances was abnormally small and broad. The other resonances were sharp and of normal intensity. It was found that these abnormal resonances occurred at points where the cavity could be simultaneously resonant in two different modes; for instance, when three half wave-lengths of the H_2 mode are equal to two half wave-lengths of the H_0 mode. Thus at a wave-length of 1·37 cm., the second H_0 response in a cavity of 19 mm. diameter was found to be smaller by a factor ten than the usual value. The adjacent H_0 resonances were, of course, unaffected. The wave-length at which n_0 half wave-lengths of H_0 are equal to n_1 half wave-lengths of another mode is given by the formula

$$\frac{\lambda}{d}=\frac{n_0^2-n_1^2}{n_0^2x_1^2-n_1^2x_0^2},$$

where d is the diameter of the barrel, and x_0 and x_1 are the values of d/λ_c appropriate to H_0 and the other mode respectively (see table 1). Some of the more important points where resonances coincide are given in the following table:—

Table 2

Interfering type	n_0	n_1	λ/d
E_2	1	0	0·612
H_1	1	2	0·733
E_0	1	2	0·749
H_3	2	1	0·730
H_1	2	3	0·644
E_0	2	3	0·672
H_2	2	3	0·722

Since the resonances due to other modes are too small to be detected by themselves, their effect on the H_0 mode at coincidence is presumably due to some asymmetry in the cavity which causes a transfer of power from the H_0 to the undesired mode. An obvious asymmetry is the presence of the coupling holes which disturb the normal current flow within the resonator. Evidence supporting this suggestion was obtained from a "reaction" wave-meter, in which the output coupling was absent and resonance was detected by the absorption of part of the power transmitted along the feeder wave-guide. The only "crossover" point at which the response was low was for two half wave-lengths, H_0 mode, equal to three half wave-lengths, H_2 mode. When a similar experiment was tried with a cavity containing an output hole, the effect of crossover with the H_1 mode could also be observed.

So far as could be ascertained, no shift of the position of resonance occurred at a crossover, and the performance of the cavity as a wave-meter was not affected. The presence of crossovers may, however, cause considerable errors if the cavity is being used to measure absorption or power factor. This is illustrated by an effect observed during use of an evacuated cavity at a wave-length where a crossover occurred, and the response was about ten times smaller than normal. When air was admitted to the cavity the response rose by a factor of three. The slight change in dielectric constant removed the coincidence of the resonances, owing to the different dispersion of the two modes. The difficulties associated with the presence of crossovers make it advisable, in absorption measurements, to use a cavity of the smallest possible diameter and to work on the first or second half wave-length, where crossovers occur infrequently. A check on the presence of a crossover should always be made by observation of several resonant positions.

The effect of these mode crossings may be reduced by damping the unwanted mode. This damping must be effected without introducing more discontinuities in the cavity, since further asymmetry is likely to increase the coupling between the modes and thus enhance rather than diminish the effect of crossovers. A good method utilizes the leakage of all modes except the H_0 past the non-constant plunger; a ring of an absorbent material such as bakelite-bonded paper or linen placed on the back of the plunger is quite effective. In a cavity designed for vacuum work (Bleaney and Penrose, 1946) water contained in an annular glass ring has also been used.

PART II.—DIELECTRIC PROPERTIES OF SIX NON-POLAR LIQUIDS AT WAVE-LENGTHS OF 3·2 CM. AND 1·35 CM.

§ 6. INTRODUCTION

The use of cavity resonators for measuring the dielectric properties of solids and liquids is well known. Penrose (1946) and others have used a cavity excited in the H_0 mode for measurements on solid dielectrics. A specimen of known dimensions is placed in the cavity, and from the shift of the position of resonance and the breadth of the resonance curve the dielectric constant and the power factor can be calculated. A similar method has been used (Whiffen and Thompson, 1946; Horner, Taylor, Dunsmuir, Lamb and Jackson, 1946) for liquids of high power factor. There are, however, several disadvantages in using a cavity only partly filled with liquid:

1. The wave-length in the liquid is not measured directly but deduced from the shift of the resonances in the air-filled portion of the cavity.

2. The height of the liquid in the cavity must be known accurately.

3. The surface of the liquid is not flat because of surface-tension effects.

4. A minor objection is that the remainder of the cavity is filled with the vapour. The error in the dielectric constant arising from this source will be only a few parts in 10 000.

For measurements on liquids with good power factors, these disadvantages may be eliminated by using a cavity completely filled with the liquid. Both the experimental procedure and the mathematics are simplified: thus, for the dielectric constant only two measurements of wave-length are required after the preliminary determination of the diameter of the cavity. In addition, the broadening of the resonance curve due to the presence of an imperfect dielectric is, of course, greatest when the cavity is filled with the dielectric.

The dielectric constant may be calculated from measurements on a filled cavity by means of the simple formula:

$$\left.\begin{array}{ll} \text{Air filled cavity:} & \dfrac{\epsilon_a}{\lambda^2}=\dfrac{1}{\lambda_a{}^2}+\dfrac{1}{\lambda_c{}^2}. \\[2ex] \text{Liquid filled cavity:} & \dfrac{\epsilon}{\lambda^2}=\dfrac{1}{\lambda_d{}^2}+\dfrac{1}{\lambda_c{}^2}, \end{array}\right\} \quad \ldots\ldots(6.1)$$

where λ = wave-length in free space,

λ_a = ,, ,, in air filled resonator,

λ_d = ,, ,, in liquid filled resonator,

λ_c = critical wave-length of cavity = diameter/1·2197,

ϵ_a = dielectric constant of air = 1·0006,

ϵ = dielectric constant of liquid.

§ 7. THE APPARATUS

Resonator A, for 3·2 cm. wave-length (barrel diameter 5·075 ± 0·001 cm.) had a side coupling to the detector, as described in § 3. The cavity was made liquid-tight by clamping thin discs of mica over the coupling holes.

A side output has the drawback that its electrical position in the resonator varies with the dielectric constant of the filling. This defect was avoided at 1·35 cm. wave-length by using an output from the end, as described in § 4; this arrangement has also the advantage of requiring only one piece of mica to close all the coupling holes. Resonator B had a diameter of $2{\cdot}106_5 \pm 0{\cdot}001$ cm., and to fit the output on to the end, a rectangular guide smaller than the normal was used. This was filled with ebonite to increase its cut-off wave-length beyond 1·35 cm. In addition, the precision wave-meter of § 4 (here called resonator "C") was used, as its large diameter makes negligible the error due to uncertainty in the barrel diameter.

When the cavity is filled with liquid, the number of possible modes of resonance is increased. The higher modes leak past the plunger, and can therefore be damped down by placing an absorbent material behind the plunger. A sealed glass tube containing water was used thus in resonator A, as most other "lossy" materials would contaminate the organic liquids under investigation; this proved unnecessary in the two smaller resonators with end output, which seems to be superior in this respect to a side output. The probability of coincident resonances is also increased by the dielectric constant of the liquid, but they could be avoided by slight readjustment of the oscillator frequency. At the shorter wave-length, four resonant points could be obtained in the liquid, and any resonance abnormally broad and weak was discarded.

Drift of the oscillator frequency was negligible at 3·2 cm. wave-length. At 1·35 cm. wave-length a small correction was necessary, determined by successive measurements with the cavity alternately filled with air and with liquid.

§8. RESULTS FOR THE DIELECTRIC CONSTANT

The values for the dielectric constants of the six non-polar liquids at a temperature of 20° c. are shown in table 3.

Table 3. Dielectric constant at 20° c.

Liquid	Purity	Dielectric constant			
		λ=3·2 cm.	λ=1·35 cm.		Low-frequency values
		Resonator "A"	Resonator "B"	Resonator "C"	
Benzene	Commercial	2·2850	2·2853		
„	Analar	2·2835	2·2828	2·2830	
„	Analar dried over sodium	2·2780	2·2778	2·2776	2·2825 (*a*)
Cyclo-hexane	"A"	2·0244	2·0246	2·0251	2·020 (*b*)
n-heptane	"A"	1·9220	1·9223	—	1·926 (*c*)
CS_2	"A"	2·6476	2·6477	—	2·630 (*d*)
n-hexane	Commercial	1·9016	1·9016	—	1·919 (*e*)
CCl_4	Commercial	2·2386	2·2390	—	2·2409 (*f*) 2·2445 (*g*)

Notes

"A" Purified sample kindly lent by Dr. H. W. Thompson of the Physical Chemistry Laboratory, Oxford.

(*a*) Hartshorn and Oliver, 1929. *Proc. Roy. Soc.*, **123**, 664. Values ranging from 2·276 to 2·289 have been obtained by different experimenters (see reference (*g*) below).

(*b*) Hooper and Kraus, 1934. *J. Amer. Chem. Soc.*, **46**, 2265.

(*c*) Smith, Morgan and Boyce, 1928. *J. Amer. Chem. Soc.*, **50**, 1883.

(*d*) Wolfke and Mazur, 1932. *Z. Phys.*, **74**, 110.

(*e*) Chrétien, 1931. *C.R. Acad. Sci., Paris*, **192**, 1385.

(*f*) Goss, 1933. *J. Chem. Soc.*, p. 1341.

(*g*) Clay, Dekker and Hemelrijk, 1943. *Physica*, **10**, 768.

In no case do the values obtained in the different resonators differ by more than 0·03%. This is consistent with the possible errors, which are as follows:—

(1) *Error in the barrel diameter*

Variation of the diameter and uncertainty in its measurement amounted to 0·001 cm. From equations (6.1) one finds

$$\frac{\delta\epsilon}{\epsilon}=2\left(\frac{\lambda}{\lambda_c}\right)^2\left(\frac{\epsilon-1}{\epsilon}\right)\frac{\delta\lambda_c}{\lambda_c}. \qquad \ldots\ldots(8.1)$$

The corresponding errors in the dielectric constant are 0·01% in resonators "A" and "C" and 0·02% in "B".

(2) *Periodic error in the micrometer*

Calibration by the N.P.L. showed that this was not more than 0·0002 mm. The scale was normally read to half a division, or 0·001 mm., and an error of this magnitude corresponds to less than 0·01% in ϵ.

(3) ***Frequency drift of the oscillator***

In the worse case ($\lambda = 1{\cdot}35$ cm.) the percentage change in frequency was not more than 0·03% during the experiments on any one liquid, corresponding to 0·02% in ϵ. A correction has been applied as described above.

(4) ***Change in temperature***

An error of 0°·2 c. in the temperature would cause an error in ϵ of 0·01%. The resonators were lagged and the temperature measured with a probable error of 0°·2 with a thermometer previously compared with a standard. The liquid in the resonator could be stirred effectively by spinning the plunger rapidly.

§ 9. THE TEMPERATURE COEFFICIENT OF THE DIELECTRIC CONSTANT

By cooling or heating the resonator slightly, the dielectric constant was determined at a number of temperatures within the range (6–30° c.). The temperature coefficient was found from a plot of ϵ against t. A correction was applied for the expansion of the metal parts of the resonator.

The values of $(d\epsilon/dt)_{20^\circ}$ obtained at a wave-length of 3·2 cm. are shown in column (2) of table 4, together with the low-frequency values given in the International Critical Tables. The variation of the dielectric constant with the temperature is principally due to expansion of the liquid. From the Clausius–Mossotti formula one finds

$$\frac{d\epsilon}{dt} = -\frac{(\epsilon-1)(\epsilon+2)}{3}\cdot\gamma,$$

where γ is the coefficient of cubical expansion, assuming that the molar polarizability is independent of temperature. Using the expansion coefficients given in the Landolt–Börnstein tables, the values of $d\epsilon/dt$ have been calculated and are shown in the last column of table 4. They are generally in good agreement with the measured values.

Table 4

Liquid	$\frac{d\epsilon}{dt}$ at 20° c.	$\frac{d\epsilon}{dt}$ at 20° c. (according to I.C. Tables)	$\frac{d\epsilon}{dt}$ calculated from expansion coefficient
Benzene—Commercial	$-{\cdot}0020 \pm {\cdot}0001$	$-{\cdot}00205$	—
„ —Analar	$-{\cdot}0020 \pm {\cdot}0001$	$-{\cdot}00198 \pm {\cdot}00003$	·0020
***n*-hexane**	$-{\cdot}0015 \pm {\cdot}0001$	$-{\cdot}0011$	·0016
Cyclo-hexane	$-{\cdot}0016 \pm {\cdot}0001$	$-{\cdot}0016$	·0016
***n*-heptane**	$-{\cdot}0013 \pm {\cdot}0001$	$-{\cdot}0014$	·0015
Carbon tetrachloride	$-{\cdot}0019 \pm {\cdot}0001$	$-{\cdot}0014$	·0021
Carbon bisulphide	$-{\cdot}0026 \pm {\cdot}0001$	—	·0030

§ 10. THE POWER FACTOR OF THE LIQUIDS

By measurement of the width of the resonance curve of the cavity when filled with liquid, a maximum value of the power factor of the liquid can be obtained. To find the true value it is necessary to correct for the dissipation of energy in the cavity elsewhere than in the dielectric. These other losses are due to the finite conductivity of the walls of the cavity and to the finite size of the coupling holes by which energy is fed into and out of the cavity. By keeping these holes small, it is generally possible to make these coupled losses small (e.g. 10%) compared with the loss of energy on the walls. The latter cannot be measured directly when the liquid is present: it might be calculated assuming the conductivity of the wall to be given by the low-frequency conductivity of the bulk metal. This will certainly give a minimum value for the loss on the walls; in practice the losses are greater than this theoretical value, as is shown by the fact that the measured Q values of empty cavities are generally found to be somewhat lower than the theoretical value at centimetre wave-lengths.

These difficulties in making a correct allowance for losses other than that in the dielectric emphazise the importance of making these losses as small as possible in comparison with that in the dielectric. In a cavity resonant in the H_0 mode, the losses on the wall are small; the correction to the power factor is about 0·0001 in these experiments. From the measured Q value of the empty cavity, an apparent conductivity of the walls may be calculated, assuming the conductivity to be the same everywhere. This apparent conductivity is then used to calculate the energy lost on the walls when the liquid is present, since the current distribution is known once the dielectric constant has been determined. This method is valid when the coupled loss is small compared with the loss on the walls; it is likely to be more inaccurate when the liquid under test has a high dielectric constant, since the change in the current distribution will be large. In these experiments the dielectric constants of the liquids were all about 2, and the coupled loss was about 5% of the wall losses; the correction should therefore be reliable.

The values of the power factors of the liquids at two wave-lengths, 3·2 cm. and 1·35 cm., are shown in table 5. The correction for the losses on the walls of the cavity has been applied; it amounted to $0{\cdot}0001_0$ at 3·2 cm., and $0{\cdot}0001_3$ to $0{\cdot}0001_8$ (depending on the dielectric constant) at 1·35 cm.

Table 5. Power factors of the liquids

Liquid	Purity	Power factor $\lambda=3{\cdot}2$ cm.	Power factor $\lambda=1{\cdot}35$ cm.
Benzene	Commercial	0·00057	0·0017
,,	Analar	0·00050	0·0012
,,	Analar dried over sodium	0·00035	0·00087
Cyclo-hexane	" A "	0·00005	0·00019
n-heptane	" A "	0·00037	0·00076
CS_2	" A "	0·00024	0·00072
n-hexane	Commercial	0·00034	0·00076
CCl_4	Commercial	0·00031	0·00078

§ 11. CONCLUSION

The internal consistency of the measurements of the dielectric constant is very gratifying, and accords with the accuracy expected of the method. The apparatus is simple, and only metrological measurements are required. The differences between these values and those previously obtained by other experimenters at low frequencies are generally somewhat greater than the possible error in the measurements. There is no reason to expect that the dielectric constants at centimetre wave-lengths of these non-polar liquids should be appreciably different from the static values, and the difference may be ascribed to small quantities of impurity. Thus the value for the purified benzene is about 0·2% lower than the generally accepted value of Hartshorn and Oliver (1929), but lies within the range of low-frequency values obtained by various methods between 1922 and 1943.

All the liquids measured show appreciable power factors at these short wave-lengths, being greater by a factor of 2 to 3 at 1·35 cm. than at 3·2 cm. This may be due to polar impurities; from the measurements of Whiffen and Thompson (1946) the necessary concentration of polar impurity can be estimated as a few hundredths of one per cent. The presence of such small amounts cannot be excluded, though it is surprising that the power factors of the pure samples "A" and the dried analar benzene are no better than those of the samples of CCl_4 and *n*-hexane, which were ordinary commercial grade. Only one liquid, cyclo-hexane, was appreciably better in power factor than the rest; for this liquid the measured power factors scarcely exceed the possible experimental error.

ACKNOWLEDGMENT

The work described in Part I of this paper was carried out under research contract for the Director of Scientific Research, Admiralty, and the authors wish to thank the Board of Admiralty for permission to publish the paper.

REFERENCES (see also notes to table 3)

Bleaney and Penrose, 1946. *Proc. Roy. Soc.* (In publication.)
Penrose, 1946. *Trans. Faraday Soc.* (In publication.)
Whiffen and Thompson, 1946. *Trans. Faraday Soc.* (In publication.)
Horner, Taylor, Dunsmuir, Lamb and Jackson, 1946. *J. Inst. Elect. Engrs.*, **93**, Part III, 53.

THE BEHAVIOUR OF WATER UNDER HYDROSTATIC TENSION: III

By H. N. V. TEMPERLEY,

King's College, Cambridge

MS. received 9 *August* 1946

ABSTRACT. **Some further experiments on the behaviour of water under tension are described, which appear to confirm the conclusions of the first two papers in this series. It appears that, under favourable conditions, water in a glass tube can support tensions as high as 60 atmospheres. A simple theoretical investigation shows that the commonly held view that the tensile strength of a liquid should be numerically equal to the "intrinsic pressure" is false. For water, a reasonable theoretical value would be 500 to 1000 atmospheres for the tensile strength, which is higher than anything that has been actually measured, though two possible explanations of this discrepancy can be suggested. In any case, it is clear that the discrepancy is much less serious than is usually supposed.**

The same theoretical considerations can be applied to the experiments of Kenrick, Gilbert and Wismer on the superheating of liquids, and their results can be accounted for without any extra assumptions.

§1. INTRODUCTION

In two previous papers, Temperley and Chambers (1946 a), referred to as Part I, and Temperley (1946 b), referred to as Part II, some measurements of the critical tensile strength of water were described. The Berthelot method of heating water in a sealed tube until the tube filled, and then cooling it until failure occurred, was examined in detail, and evidence was obtained that quite high pressures might be developed in the tube before the final disappearance of the gas bubble by solution in the water. The assumption of zero pressure at this instant thus leads to spuriously high readings for the tension in the liquid when failure occurs. In Part I it was suggested that the high pressures were necessary to force the water to enter fissures in the glass, but in Part II clear evidence was obtained that a small gas bubble can exist for long periods in the presence of water at high pressure. A simple calculation showed that the observed rates of diffusion of gases in liquids are so low that a gas bubble would have to be compressed to a diameter of a few tenths of a millimetre to make it dissolve in a reasonable time, so that one can account for the observed high pressures in this manner also. In an attempt to decide between these two hypotheses, some experiments were carried out with a wetting agent added to the water, the original idea of which was to facilitate the flow of water into fissures in the glass. A second effect of the wetting agent was to prevent the very small bubbles that are formed at the instant of failure from joining up again immediately, as they do in plain water. The experiments described below appear to provide decisive evidence that the high pressures are necessitated by the slowness of diffusion.

An attempt, only partially successful, was made to reproduce the mechanism of the ejection of spores from ferns (King, 1944). The mechanism is the diffusion of water vapour through the cell wall, with a consequent shrinkage of the cell and the setting up of tension in the remaining liquid, the cell springing back to its normal shape when the liquid breaks. Use was made of the observation of Reekie and Aird (1945) that a tube filled with wet jeweller's rouge is permeable to water vapour but not to air.

The paper ends with a theoretical discussion of the problem of the strength of liquids. Quite simple considerations show that the assumption made by almost all writers (including the author in Parts I and II), that the so-called "intrinsic pressure" of a liquid is the same thing as its theoretical tensile strength, is false. The latter quantity is very much the smaller, though still larger than any tension that has actually been measured in water. There is a little evidence that it may be possible to approach the theoretical value by pre-compression of the water to destroy possible nuclei. The same considerations account nicely for the results obtained by Kenrick, Gilbert and Wismer (1924) on the superheating of liquids.

§2. EXPERIMENTS WITH WATER CONTAINING A WETTING AGENT

The wetting agent used was sulphonated lauryl alcohol, and I am extremely grateful to Dr. A. E. Alexander for it. It was used in approximately 1% solution, and a rough measurement (by the capillary rise method) gave a value of 25 dynes/cm. for the surface tension. Some Berthelot tubes were made up, filled with this solution and the tensile strength measured by the two methods described in Part II (the dilatometer and photo-elastic methods). Again there was satisfactory agreement between the two methods within the very limited accuracy of the photo-elastic method, but no appreciable difference between these results and those for pure water could be found, a total of twelve experiments on three different tubes giving results ranging between 20 and 60 atmospheres true tension. Mean 32 atmospheres. (A correction was made for the calorimetric effect as before.) However, a very considerable difference in the behaviour of the gas bubbles during the heating of the tube was observed, depending on whether they were large or small. At the moment of failure a multitude of very small bubbles, radii about 0·1 mm., appeared. If the tube was now immediately reheated, these tiny bubbles dissolved almost immediately, and it was found that the process could be further expedited by inverting the tube a few times, thus bringing the bubbles constantly into contact with fresh water. If, however, the tube was allowed to stand for an hour or more, the small bubbles joined up to form large ones, and the behaviour of the tube was then indistinguishable from that of one containing plain water.

The following example is typical : the filling temperature of one tube was determined with great care by the method described in Part II, the bath being heated very cautiously so that the process of dissolving the gas bubble occupied about an hour. This temperature was found to be 52° c. and the breaking temperature was determined to be $34 \pm \frac{1}{2}°$ c. If, however, the tube was reheated immediately after failure of the liquid, while the bubbles were still small, it was found possible to refill it in a few minutes, even though the temperature of the bath was no more than 42° c. The exact pressure in the tube at this temperature is uncertain, but the absence of any photo-elastic effect shows that it is less than 10 atmospheres, whereas it seems to require not less than 50 atmospheres to cause a single large bubble to dissolve in a reasonable time. The breaking temperature at which the water failed under tension was the same, whichever method of filling was used. It therefore seems certain that the slowness of diffusion is responsible for the high pressures that occur during the filling of Berthelot tubes, and also that any effect of pressures, of the order of 50 to 100 atmospheres, forcing water into fissures in the glass has little effect on the tension that can subsequently be developed in the tube. Vincent's (1943) experiments on mineral oil also show that the pre-application of pressure of 100 atmospheres has little effect.

§ 3. AN ATTEMPT TO IMITATE THE MECHANISM OF SPORE-DISCHARGE

A tube was sealed at one end and a constriction formed near the other end. The tube was filled with water that had been well boiled to remove as much air as possible, and about 1 cm. of the tube above the constriction was filled with jeweller's rouge packed as tightly as possible. The open end was then connected to a calcium-chloride tube and the latter to a filter pump. At equilibrium the pressure of water vapour near the calcium chloride would be negligibly small, so one might expect tension to be developed in the water corresponding to the pressure difference that can be supported by surface tension acting in the small channels between the particles of rouge. It was shown by Reekie and Aird (1945) that such a wet pad of rouge could withstand air pressures of at least three atmospheres, but was readily permeable to water vapour. We thus seem to have a fair imitation of the mechanism of transpiration in plants.

The results of the experiments may be stated quite shortly. It was not found possible to cause pure water to fail by tension by this method, which is hardly surprising, because a tension of 40 atmospheres corresponds to an effective channel radius of only $3{\cdot}5 \times 10^{-6}$ cm. It was, however, possible to produce failure with a slight "click" if a drop of mercury had been added to the tube. The tension required to rupture a water-mercury interface is uncertain. Some experiments with Berthelot tubes containing varying quantities of mercury as well as water seemed to show that the required tension was finite, because failure occurred with formation of bubbles and an audible click, exactly as in the experiments described in Part I, when pieces of steel were introduced into the tubes, but attempts to measure this tension were not successful, which probably means that it was less than 10 atmospheres.

§ 4. THEORETICAL INVESTIGATION OF THE TENSILE STRENGTH OF LIQUIDS

We consider a liquid that obeys the van der Waals equation of state. It is known (see Mayer and Mayer, *Statistical Mechanics*) that almost any law of force between the molecules leads to the van der Waals equation as a first approximation. It would not be difficult to modify the ensuing theory if a more accurate equation of state were known for a particular liquid. We write the equation of state in the form

$$(P + a/V^2)(V - b) = RT. \qquad \ldots\ldots(1)$$

It is well known that this equation predicts a discontinuous change of state if there is any region for which $(\partial P/\partial V)_T$ is positive, such as the portion AB in figure 1, according to equation (1), for any point on this portion of the curve the thermodynamic potential takes a stationary value; but it is a maximum, not a minimum, and this portion of the curve is thus not realized in practice. The pressure at which the change of state from liquid to vapour in equilibrium with it takes place is determined by the well known rule of equal areas (dotted lines in figure 1). This rule is a simple consequence of the fact that, at equilibrium, the thermodynamic potentials of liquid and vapour must be equal, and of the thermodynamic relation $(\partial G/\partial P)_T = V$. Below this equilibrium vapour pressure, the liquid is still capable of existing in a metastable state, provided that $(\partial P/\partial V)_T$ is still negative according to equation (1). The thermodynamic potential will still be a minimum with respect to small alterations of the volume, but it will no longer be an ***absolute*** minimum. It follows that the limit of the metastable region is determined by the condition that $(\partial P/\partial V)_T$ vanishes, e.g. points such as A, C, and O in figure 1, because then the thermodynamic potential no longer assumes even a local minimum, but has instead a point of inflexion, so that a transition to the vapour state will certainly be able to occur without even a temporary increase in the thermodynamic potential during the process, metastable state being impossible any longer.

Thus, by differentiating equation (1) with respect to V, we find that the absolute limit of the metastable region is given by the condition

$$P - a/V^2 + 2ab/V^3 = 0. \qquad \ldots\ldots(2)$$

The corresponding pressure will be positive or negative according as V is greater or less than $2b$. At the critical temperature $(\partial^2 P/\partial V^2)_T$ vanishes as well as $(\partial P/\partial V)_T$ and we derive the well known relations

$$V_c = 3b, \qquad P_c = a/27b^2, \qquad RT_c = 8a/27b. \qquad \ldots\ldots(3)$$

If the liquid is just capable of existing in the metastable state at zero pressure (point C in figure 1), equation (2) shows that we must have $V = 2b$, and substitution in equation (1) gives us $RT_m = a/4b$ or

$$T_m = 27T_c/32 \qquad \ldots\ldots(4)$$

for the corresponding temperature. Below this temperature, the region of metastability should extend to negative pressures, and the liquid should accordingly be capable of standing tension (point D in figure 1).

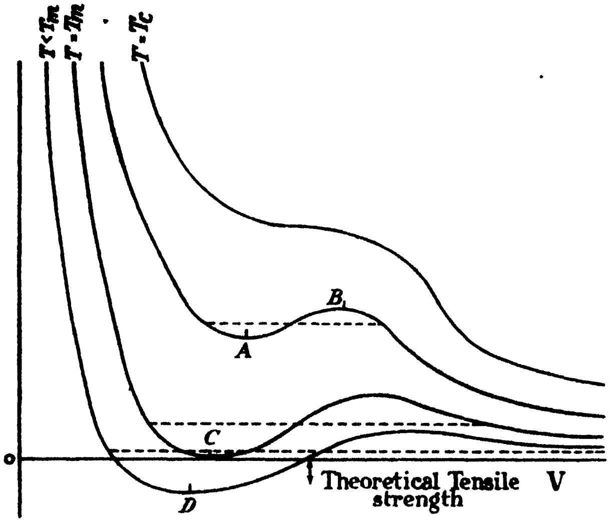


Figure 1. Isothermals of a van de Waals gas. Dotted lines indicate equilibrium vapour pressure.

Equation (4) predicts theoretically the highest temperature at which the liquid state can exist at all at zero pressure. The superheating of a number of liquids at atmospheric pressure was studied by Kenrick, Gilbert and Wismer (1924), who found that there was a very definite limiting temperature for each liquid, at or below which it always exploded. Their values for these temperatures are plotted in figure 2, and it will be seen that there is good agreement with equation (4), even with the rather crude approximation implied by the van der Waals equation.

Equation (2) enables us to make a theoretical estimate of the behaviour of the tensile strength of a liquid as a function of temperature. The tensile strength is clearly zero at $T = T_m$ where $V = 2b$. At absolute zero, supposing the equations were still valid, we should have $V = b$, and equation (2) then gives $P = -a/b^2$, that is, the tensile strength would be 27 times the critical pressure. (Only near the absolute zero is the tensile strength comparable with the "intrinsic pressure" a/V^2.) By giving V various values between $2b$ and b, we can, by using equations (1) and (2), determine the tensile strength as a function of temperature. Because of the law of corresponding states, the theoretical curve is the same for all liquids except for scaling factors. It is plotted in figure 3 in non-dimensional form, though the positions on the curve that would be occupied by the boiling and freezing points of water have been indicated. For water, we have $T_m = 270°$ c.

It will be seen that the theoretical tensile strength of liquid water may be expected to vary between 0·1 and 0·2 a/b^2, that is, between 500 and 1000 atmospheres if we take the value of a/b^2 deduced from the critical pressure. There is no actual numerical evidence of tensions as high as this, though it is clear that the discrepancy between theory and experiment is very much less than is usually supposed. Moreover, evidence that the tensile strength of water can be increased by pre-compression or strong centrifuging has already been obtained by Harvey and others (1944.) (In Part I it was noted that no method of completely freeing water from dissolved gases had been found.) The point that appears to have been overlooked is that, at any finite temperature, the tension necessary to cause failure in a liquid may be numerically very much less than the "intrinsic pressure" a/V^2. It should be mentioned that there is an alternative method of estimating the constant a, using the latent of heat of evaporation. If the van der Waals equation held for all volumes, the work done in increasing the volume from V to infinity would be

$$RT\log(V-b)+a/V,$$

the term a/V representing the work done against the attraction of the molecules. If we identify this with the latent heat of evaporation, we get a value of

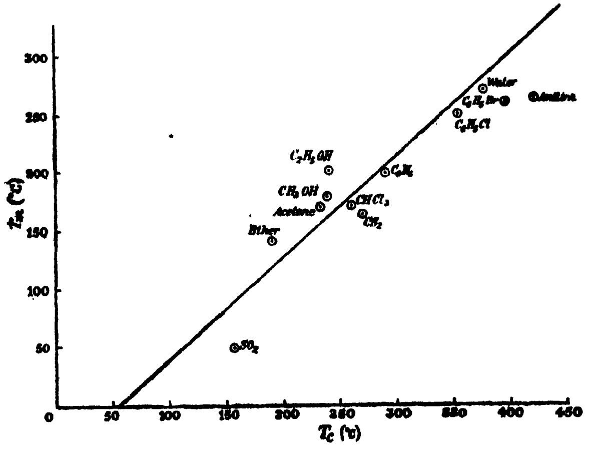


Figure 2. Comparison with theory of the results of Kenrick, Gilbert and Wismer on the superheating of liquids.

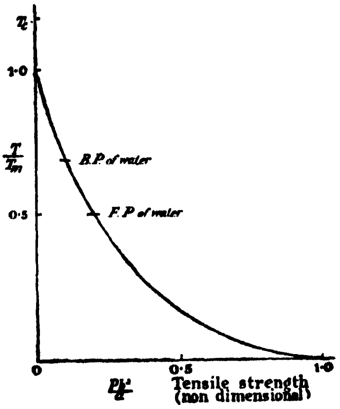


Figure 3. Theoretical tensile strength of van der Waals' liquid as a function of absolute temperature.

T_m = Limiting temperature for exsitence of liquid at atmospheric pressure.
T_c = Initial temperature.
Straight line $\frac{T_m}{T_c} = \frac{27}{32}$ deduced from van der Waals' equation.

25 000 atmospheres for the intrinsic pressure and correspondingly larger values for the tensile strength. However, this argument involves a long extrapolation, and it is known that water does not obey van der Waals' equation very well. The former argument involves the assumption that the constants a and b are independent of temperature so that they may be calculated from the critical data, and is thus also not free from objection. The theoretical tensile strength is sensitive to the value of b, as well as to that of a, but it is interesting that one can account for the experimental facts on superheating of liquids *without* assuming that these quantities vary with temperature, in fact, relation (4) is not much affected by such variations. It seems worth while to ask two further questions. First, what is the effect on the theoretical predictions of assuming the quantities a and b to be dependent on temperature; secondly, what can be said theoretically about the probable behaviour of a and b as functions of temperature?

§ 5. POSSIBILITY OF VARIATIONS OF THE VAN DER WAALS CONSTANTS WITH TEMPERATURE

The discussion of the effect of possible variations of b with temperature is quite simple. It can easily be shown from equations (1) and (2) that if b increases above the value $V_c/3$, the tensile strength at a given temperature will be smaller than if b remains constant. Now, since b is a measure of the "excluded volume" it is fairly obvious that b must decrease as the temperature rises and the collisions become more energetic, except in the limiting case of rigid molecules for which it should remain constant. As the temperature falls, b should increase, so that we might expect the tensile strength to be less than the value predicted by equations (1) and (2) from the critical data. It is also fairly clear that an increase of a, which is a measure of the attraction between the molecules, should increase the predicted tensile strength, but it is not so simple to decide whether a should increase or decrease with increasing temperature.

The following rough argument seems to show that either type of behaviour is possible. Let us suppose that we specify the state of the liquid by means of a variable r, which measures the average distance apart of neighbouring molecules. As a rough approximation, the free volume occupied by N molecules may be taken as KNr^3, where K is some constant depending on the shape of the molecules and the average number of nearest neighbours. Thus there will be a contribution to the entropy of the order of magnitude $\boldsymbol{R}\log(KNr^3)$, corresponding to the term $\boldsymbol{R}\log V$ in a perfect gas, arising from the various possible ways of arranging the molecules. Now suppose we assume that the mutual potential energy of two molecules can be expressed by means of an attraction and a repulsion, both of power-law type. The partition function will then be given by an expression of the type, qua function of r:

$$\log f = (\text{terms indep. of } r) + N\log(KNr^3) - \left(\frac{-A}{r^m} + \frac{B}{r^n}\right)\Big/ \boldsymbol{k}T. \qquad \ldots\ldots(4)$$

We are assuming that the external pressure is zero, and also that the parts of the partition function depending on the internal degrees of freedom of the molecule are not affected appreciably by changes of r. For a small external pressure, the state of absolute equilibrium is the vapour phase (r very large), but a state of metastable equilibrium is possible for small r provided that $\partial(\log f)/\partial r$ vanishes, and provided also that $\partial^2(\log f)/\partial r^2$ is negative, so that the partition function is a maximum, not a minimum. We recall that a maximum of the partition function corresponds to a minimum of the thermodynamic potential. The temperature T_m, corresponding to the limit of metastability being at zero pressure, will correspond to the vanishing of both these quantities, so that we have

$$3NkT_m = \frac{mA}{r_m^m} - \frac{nB}{r_m^n} = \frac{m(m+1)A}{r_m^m} - \frac{n(n+1)B}{r_m^n}. \qquad \ldots\ldots(5)$$

Eliminating T_m, we find

$$r_m = \left(\frac{n^2B}{m^2A}\right)^{\frac{1}{n-m}}. \qquad \ldots\ldots(6)$$

The average attraction between molecules is proportional to the derivative of the potential energy with respect to r, and this will be a decreasing or increasing function of r according to the sign of the second derivative. The value of r for which the average attraction is a stationary function of r is given by

$$r_1 = \left(\frac{n(n+1)B}{m(m+1)A}\right)^{\frac{1}{n-m}}. \qquad \ldots\ldots(7)$$

If there were no thermal agitation, the value of r would be given by

$$r_0 = \left(\frac{nB}{mA}\right)^{\frac{1}{n-m}}. \qquad \ldots\ldots(8)$$

If we give m and n reasonable values, say 6 and 12 respectively, we see that r_1 is considerably greater than r_0, but only about 1% less than r_m. Since the volume is approximately proportional to r^3, it follows that r_1 corresponds to a temperature of about 50°–100°c. below T_m if we assume a reasonable value of 5 to 10×10^{-4} for the coefficient of expansion of the liquid per °c. Thus, it seems likely that practically all the experiments that have been done to produce tension in liquids have taken place at a temperature below that corresponding to r_1, and thus in a region where we should expect the *average* attraction between molecules, and thus also the van der Waals constant a, to be increasing functions of the temperature. Incidentally, we also seem to have obtained an explanation of the observed fact that the van der Waals equation can be made to give fair agreement with experiment over quite considerable ranges of temperatures, because b probably changes fairly slowly with temperature, and a seems to pass through a maximum at a temperature in the liquid range.

It is hardly necessary to point out that an increase of a corresponds to an increase in tensile strength of the liquid, and this can be verified by inspection of equations (1) and (2). It is, however, not possible to say whether a liquid at a temperature of, say, 30° c. should have an effective value of a greater or less than that at its critical temperature, so that we are unable to predict the nature of the departure from the curve in figure 3 due to the variation of a with temperature. The argument indicates that the modified curve would lie above the curve in figure 3 at the temperature corresponding to r_1, and that it would cross it at some lower temperature.

It thus seems likely that any "softness" of the molecules would reduce the tensile strength below the value calculated from the critical data, but it is not possible to make any numerical prediction about the variation of the constant a with temperature, though the effect of such a variation can be calculated immediately. On the other hand, one can be fairly confident about the sign of any variation of b, and of the effect of this. It seems just worth mentioning an interpretation of this effect of "softness" of molecules. An increase in "softness" increases the volume within which the centre of a molecule may be found, and thus favours the possible formation of "holes" in a liquid. In fact, an attempt has been made by Fürth (1941) to estimate the probability of the appearance of "holes" of various sizes. The theory indeed predicts a metastable state of a liquid under tension, but the results are unlikely to be of quantitative value, because the author uses, as he himself points out, concepts such as surface energy and hydrodynamic virtual mass, which are essentially macroscopic, in connection with cavities of molecular dimensions. In fact, the theory predicts tensile strengths which are much too high.

Actual calculations have shown that the theoretical values of the temperature T_m (figure 2), are much less affected by variations of a and b with temperature than is the theoretical tensile strength (figure 3). For example, the rather extreme assumption made in Berthelot's equation of state ($a \propto 1/T$) leads to an approximately ten-fold increase in the theoretical tensile strength of water compared with the assumption that a is constant, while the ratio between T_m and T_c is changed only from 27/32 to $\sqrt{27/32}$.

§6. CONCLUSIONS

The conclusion reached in Part II that slowness of diffusion is the factor that causes high pressures to occur in Berthelot tubes during the filling process has been confirmed, and evidence has also been obtained that the prior application of pressures of the order of 100 atmospheres has little effect on the tension that can subsequently be developed in the tube. Theoretical considerations, supported by the experimental work on the superheating of liquids, indicate that the discrepancy between calculated and observed strengths of liquids is not nearly as great as is usually supposed. Possible explanations of the remaining discrepancy are:

(*a*) The effect of dissolved gases and other nuclei.

(*b*) The departure of liquids from an equation of state of the van der Waals type.

ACKNOWLEDGMENTS

I wish to thank Sir Lawrence Bragg for the facilities for this investigation and for helpful discussion and advice, and also Sir Geoffrey Taylor for inspiring this investigation. I should also like to thank Dr. A. E. Alexander for supplying me with the wetting agent and for helpful discussion. Finally, I wish to thank the Leverhulme Trustees for a generous research grant.

REFERENCES

Fürth, 1941. "On the theory of the liquid state.—II." *Proc. Camb. Phil. Soc.*, **37**, 276.

Harvey, *et al*, 1944. "Bubble formation in animals." *J. Cellular and Comparative Physiol.*, **1**, 23.

Kenrick, Gilbert and Wismer, 1924. "The superheating of liquids." *J. Phys. Chem.*, **28**, 1297.

King, 1944. "The spore discharge mechanism of common ferns." *Proc. Nat. Acad. Sci., Wash.*, **30**, 155.

Reekie and Aird, 1945. "Flow of water through very narrow channels and attempts to measure thermo-mechanical effects in water." *Nature, Lond.*, **156**, 367.

Temperley and Chambers, 1946 a. "The behaviour of water under hydrostatic tension.—I." *Proc. Phys. Soc.*, **58**, 420.

Temperley, 1946 b. "The behaviour of water under hydrostatic tension.—II." *Proc. Phys. Soc.*, **58**, 436.

Vincent, 1943. "Examination of the Berthelot method of measuring tension in liquids." *Proc. Phys. Soc.*, **55**, 376.

AN ANALYSIS OF THE CONDITIONS FOR RUPTURE DUE TO GRIFFITH CRACKS

By H. A. ELLIOTT,
Bristol

Communicated by Professor N. F. Mott, F.R.S. ; *MS. received* 9 *September* 1946

ABSTRACT. The solutions for the problem of an infinite isotropic elastic solid stressed under tension T_0 and containing a single internal crack of length c on the plane $z=0$ are given in a form suitable for the computation of the stresses and displacements at all points. These are used to find the stress distribution on, and the displacements of, the plane situated $\frac{1}{2}a$ from the plane containing the crack. The normal stress σ_z on $z=\frac{1}{2}a$ (as found above) is plotted as a function $f(2u_z)$ of the normal displacement u_z, and τ_{rz} is small compared with σ_z.

A model is used in which the crack is considered to be bounded by the atoms centred on the planes $z=\pm\frac{1}{2}a$, these planes being the boundaries of two semi-infinite elastic solids. Equilibrium is maintained by postulating that an attractive force, $f(z)$, acts between the atoms of these bounding planes when they are $z+a$ apart. It is found that $f(z)$ approximates to the law of force expected from atomic considerations, and the condition for unstable equilibrium of the crack, i.e. a value $T_0{}^c$ of T_0 such that for $T_0<T_0{}^c$ the crack closes (c decreases), and for $T_0>T_0{}^c$ the crack spreads (c increases), is found. The surface energy is calculated from the results and the equilibrium condition is found in a form similar to that of Griffith. Agreement is found with the experimental results of Griffith.

In the absence of the tension T_0, the crack cannot be maintained without an inclusion to prevent closing. Possible physical models are discussed.

§ 1. INTRODUCTION

The object of this paper is to extend Griffith's theory (1921) of rupture through a more detailed consideration of the interatomic forces which resist the spread of a crack.

Griffith's theory may be expressed in the following form : suppose that a solid of Young's modulus E is subjected to a stress T_0, and that it contains a crack of length $2c$. Then the elastic energy (or that part of it which depends on c) is, per unit length,

$$-\frac{T_0^2c^2}{E}\times\pi(1-\sigma^2).$$

Griffith supposes that the total contribution to the energy which depends on c is obtained by adding the surface energy $4Sc$, where S is the surface energy per unit area. Thus we have for the energy

$$-\frac{T_0^2c^2}{E}\times\pi(1-\sigma^2)+4Sc.$$

Plotted against c, this gives a curve with a maximum at a critical value c_0; Griffith assumes that the crack will spread if c lies to the right of this maximum.

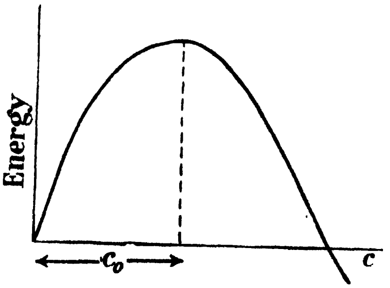


Figure 1.

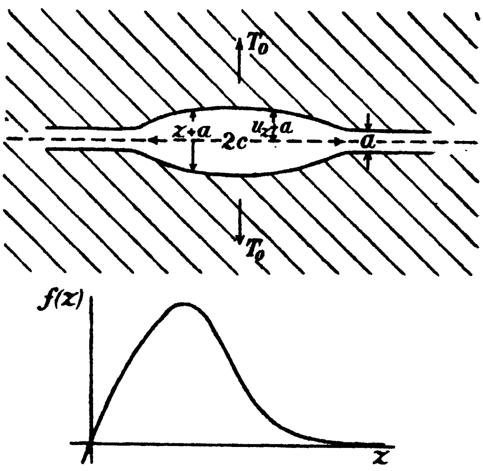


Figure 2.

In the Griffith theory it is implicitly assumed that once a crack is formed it is incapable of closing again; otherwise any crack of length less than the critical value c_0 would close. It is possible that in glass, after the silicon–oxygen bonds are broken, a displacement of the atoms in the amorphous material does in fact prevent them from reforming.

In this paper, however, we consider the fracture of a crystalline substance (e.g. a metal) along a cleavage plane. The model we take (§ 4) is of two semi-infinite blocks of solid attracting each other with a force $f(z)$ per unit area when they are at a distance $z+a$ from each other; $f(z)$ represents the force of attraction between two planes of atoms. $f(z)$ must be of the form $f(z)=Ez$ if z is small and $f(z)\to 0$ for $z\gg a$; here a denotes the interatomic distance.

The quantity S defined by $\int_0^\infty f(z)\,dz=2S$ represents the free energy of the surface per unit area.

We then consider two forms of crack in the boundary joining the two blocks: a penny-shaped crack of diameter $2c$, and a crack of width $2c$ but extending in the other direction to the boundaries of the solid, are treated. Such cracks will *either* close up *or* extend under an applied stress T_0, except for one value of c which gives unstable equilibrium. In any physical application of the theory to a material weakened by Griffith cracks we must consider some agency which holds the crack apart—e.g. adsorbed gas (possible models are discussed in § 5). In this paper, however, just as in the Griffith theory, we shall work out the condition for unstable equilibrium, i.e. the condition that the crack is about to extend (or contract). The method of approach is as follows: we first find the stress round a crack in a stressed elastic solid on the assumption that no stresses act across the surface of this mathematical crack, as in the work of Inglis (1911) and Griffith. We have actually used methods due to Sneddon (1947) and Elliott (1947) (§§ 2 and 3). We then calculate σ_z, the z component of stress, and τ_{rz} (or τ_{xz}), the shear stress, in the two planes which before stressing were distant $\pm \frac{1}{2}a$ from the plane of the crack. It is found that τ_{rz} (or τ_{xz}) is small compared with σ_z. Thus the shape of these planes is identical with that of the boundary of a crack, if the interatomic force $f(z)$ is the same function of z as σ_z is of $2u_z$ in the elastic problem (§ 4), and σ_z on these planes is now assumed to be due to interatomic forces between the planes. u_z is here the displacement in the z direction of the planes mentioned above. Near the apex of the crack, σ_z and u_z depend only on T_0, E, σ and $(c/a)^{\frac{1}{2}}$ (to the order of the approximation used), so the (σ_z, u_z) curve depends only on T_0, E, σ and $(c/a)^{\frac{1}{2}}$ for values of u_z between $0{\cdot}05a$ and a (when C is of the order of $10^4 a$).

This means that we have only the parameters, T_0, $(c/a)^{\frac{1}{2}}$, which we can use to identify our (σ_z, u_z) curve with the true interatomic force curve. In fact, we make the requirement that the maximum stress σ_z is equal to the maximum stress P obtainable from the interatomic forces. This gives us immediately a condition for rupture:—

$$T_0\sqrt{c} = \frac{4\pi}{7} P\sqrt{a} \quad (\sigma = 0{\cdot}25)$$

for the case of a penny-shaped crack, or

$$T_0\sqrt{c} = \frac{8}{7} P\sqrt{a} \quad (\sigma = 0{\cdot}25) \qquad \ldots\ldots(1.1,1)$$

for the case of a Griffith crack.

A check on the validity of this identification is given by the fact that we have now determined the surface energy S in terms of P, E and a. In fact we find,

$$\frac{S\,.\,E}{P^2\,.\,a} \approx 1 \qquad (\sigma = 0{\cdot}25). \qquad \ldots\ldots(1.1,2)$$

Approximate agreement is found in the case of the experiments of Griffith (1924).

This relationship permits the expression of our rupture condition in the same form as that of Griffith, i.e. for the penny-shaped crack and the Griffith crack respectively:

$$\left.\begin{aligned} &T_0\sqrt{c} = \frac{4\pi}{7\sqrt{2}}\sqrt{(E\,.\,S)} = 1{\cdot}27\sqrt{(E\,.\,S)} \quad (\text{for } \sigma = 0{\cdot}25) \\ \text{or}\quad &T_0\sqrt{c} = \frac{8}{7\sqrt{2}}\sqrt{(E\,.\,S)} = 0{\cdot}81\sqrt{(E\,.\,S)} \end{aligned}\right\} \qquad \ldots\ldots(1.1,3)$$

as compared with Griffith's result:

$$T_0\sqrt{c} = \sqrt{\left(\frac{2E \,.\, S}{\pi(1-\sigma^2)}\right)} \qquad \ldots\ldots(1.1,4)$$

$$= 0{\cdot}824\sqrt{(E \,.\, S)} \quad (\text{for } \sigma = 0{\cdot}25). \qquad \ldots\ldots(1.1,5)$$

§ 2. ANALYSIS OF STRESS AND STRAIN FOR A PENNY-SHAPED CRACK

2.1. In the three dimensional case, we consider a penny-shaped crack, bounded by the circle $x^2+y^2=c^2$ in the plane $z=0$. Hence the distribution of stress and displacement is the same as that for the semi-infinite elastic medium bounded by the plane $z=0$, the boundary conditions being for a crack maintained by internal pressure p_0:—

$$\left.\begin{array}{ll} \text{i.} & \tau_{xy} = \tau_{yz} = \tau_{rz} = 0 \text{ on } z=0 \text{ for all } x, y; \\ \text{ii.} & \sigma_z = -p_0 \text{ on } z=0, \quad r=\sqrt{(x^2+y^2)} \leqslant c \\ & u_z = 0 \text{ on } z=0, \quad r \geqslant c. \end{array}\right\} \qquad \ldots\ldots(2.1,1)$$

We now employ cylindrical polar co-ordinates (r, θ, z) and we use the solutions in the form given by Sneddon (1947) (equations 3.6, 2–3.6, 9):

$$u_r = +\frac{2p_0c(1+\sigma)(1-2\sigma)}{\pi E}\int_0^\infty \left(1-\frac{\zeta\eta}{1-2\sigma}\right)\frac{d}{d\eta}\left(\frac{\sin\eta}{\eta}\right)e^{-\zeta\eta}J_1(\rho\eta)\,d\eta$$

$$= +\frac{2p_0c(1+\sigma)(1-2\sigma)}{\pi E}\left[\frac{2(1-\sigma)\zeta}{1-2\sigma}S_0{}^1 - \frac{\rho}{2}S_0{}^0 + \frac{\rho}{2}S_0{}^2 - \frac{\zeta}{1-2\sigma}c_1{}^1\right], \qquad \ldots\ldots(2.1,2)$$

$$u_z = -\frac{4p_0c(1-\sigma^2)}{\pi E}\int_0^\infty \left[1+\frac{\zeta\eta}{2(1-\sigma)}\right]\frac{d}{d\eta}\left(\frac{\sin\eta}{\eta}\right)e^{-\zeta\eta}J_0(\rho\eta)\,d\eta$$

$$= \frac{4p_0c(1-\sigma^2)}{\pi E}\left[1-\frac{\zeta^2}{2(1-\sigma)}S_1{}^0 - \frac{\zeta\rho}{2(1-\sigma)}S_1{}^1 - \zeta\frac{1-2\sigma}{2(1-\sigma)}S_0{}^0 - \rho S_0{}^1\right], \qquad \ldots\ldots(2.1,3)$$

$$\sigma_z = \frac{2p_0}{\pi}[C_1{}^0 - S_0{}^0 + \zeta C_2{}^0 - \zeta S_1{}^0], \qquad \ldots\ldots(2.1,4)$$

$$\tau_{rz} = \frac{2p_0}{\pi}[C_2{}^1 - S_1{}^1], \qquad \ldots\ldots(2.1,5)$$

$$\sigma_r + \sigma_\theta + \sigma_z = \frac{4(1+\sigma)p_0}{\pi}[C_1{}^0 - S_0{}^0], \qquad \ldots\ldots(2.1,6)$$

$$\sigma_\theta - \sigma_r = \frac{2p_0}{\pi}[(1-2\sigma)(C_1{}^2 - S_1{}^2) - \zeta(C_2{}^2 - S_1{}^2)], \qquad \ldots\ldots(2.1,7)$$

where

$$\zeta = \frac{z}{c}, \qquad \rho = \frac{r}{c}$$

and

$$\left.\begin{aligned} Z_n^m &= C_n^m(\rho,\zeta) - iS_n^m(\rho,\zeta) \\ &= \int_0^\infty \eta^{n-1}e^{-\zeta\eta}J_m(\rho\eta)\,d\eta. \end{aligned}\right\} \qquad \ldots\ldots(2.1,8)$$

If we now write

$$r=1+\zeta^2, \qquad \zeta\tan\theta=1, \qquad \ldots\ldots(2.1,9)$$

$$R^2=(\rho^2+\zeta^2-1)^2+4\zeta^2, \qquad 2\zeta\cot\phi=\rho^2+\zeta^2-1,$$

we have

$$\left.\begin{aligned} C_1{}^0&=R^{-\frac{1}{2}}\cos\tfrac{1}{2}\phi,\\ C_2{}^0&=rR^{-\frac{3}{2}}\cos(\tfrac{3}{2}\phi-\theta),\\ C_2{}^1&=\rho R^{-\frac{3}{2}}\cos\tfrac{3}{2}\phi,\\ S_0{}^0&=\tan^{-1}\frac{R^{\frac{1}{2}}\sin\frac{1}{2}\phi+r\sin\theta}{R^{\frac{1}{2}}\cos\frac{1}{2}\phi+r\cos\theta},\\ S_1{}^0&=R^{-\frac{1}{2}}\sin\tfrac{1}{2}\phi,\\ S_0{}^1&=\frac{1}{\rho}(1-R^{\frac{1}{2}}\sin\tfrac{1}{2}\phi),\\ \text{and}\qquad S_1{}^1&=\frac{r}{\rho}R^{-\frac{1}{2}}\sin(\theta-\tfrac{1}{2}\phi). \end{aligned}\right\} \qquad \ldots\ldots(2.1,10)$$

2.2. The equations given in § 2.1 give the complete system of stress and displacement at all points in the medium for a crack maintained by an internal pressure p_0 in a body free from stress at infinity. For the case of Griffith's fracture we are concerned with a body under surface forces, the faces of any crack being assumed free surfaces in the elastic-theory solutions.

If these surface forces are given by tensions (or pressures), the solutions are easily found from those in 2.1. Thus for $T_z=T_0$, $T_r=T_1$ we have

$$u_r=\frac{T_0}{E}\left\{\frac{2(1+\sigma)(1-2\sigma)c}{\pi E}\left[\frac{2(1-\sigma)\zeta}{1-2\sigma}S_0{}^1-\frac{\rho}{2}S_0{}^0\right.\right.$$
$$\left.\left.+\frac{\rho}{2}S_0{}^2-\frac{\zeta}{1-2\sigma}C_1{}^1\right]-\sigma r\right\}+\frac{T_1 r}{E}, \qquad \ldots\ldots(2.2,1)$$

$$u_z=\frac{T_0}{E}\left\{z+\frac{4(1-\sigma^2)c}{\pi}\left[1-\frac{\zeta^2}{2(1-\sigma)}S_1{}^0-\frac{\zeta\rho}{2(1-\sigma)}S_1{}^1\right.\right.$$
$$\left.\left.-\zeta\frac{1-2\sigma}{2(1-\sigma)}S_0{}^0-\rho S_0{}^1\right]\right\}-T_1\frac{\sigma z}{E}, \qquad .(2.2,2)$$

$$\sigma_z=T_0\left\{1+\frac{2}{\pi}[C_1{}^0-S_0{}^0+\zeta C_2{}^0-\zeta S_1{}^0]\right\}, \qquad \ldots\ldots(2.2,3)$$

$$\tau_{rz}=T_0\frac{2}{\pi}\zeta[C_2{}^1-S_1{}^1], \qquad \ldots(2.2,4)$$

$$\sigma_r+\sigma_\theta+\sigma_z=T_0\left\{1+\frac{4(1+\sigma)}{\pi}(C_1{}^0-S_0{}^0)\right\}+T_1, \qquad \ldots\ldots(2.2,5)$$

$$\sigma_r-\sigma_\theta=\frac{2T_0}{\pi}\{(1-2\sigma)(C_1{}^2-S_0{}^0)-\zeta(C_2{}^2-S_1{}^2)\}-T_1. \qquad \ldots\ldots(2.2,6)$$

2.3. Under these conditions the shape of the crack becomes an oblate ellipsoid

$$\frac{x^2+y^2}{c^2}+\frac{z^2}{\epsilon^2}=1, \qquad \ldots\ldots(2.3,1)$$

where

$$\epsilon=[u_z]_{r=0,\,z=0}=\frac{4(1-\sigma^2)T_0c}{\pi E}. \qquad \ldots\ldots(2.3,2)$$

More exactly stated, this is the shape of the boundary found by the methods of elastic theory. However, in the actual physical case we are more interested in the displacement of the first plane of atomic centres on either side of the cleavage plane, and these lie at a finite distance, $\frac{1}{2}a$ (a = lattice spacing) from the theoretical boundary, $z = 0$.

We know from the Griffith theory (Griffith, 1921) that in normal materials the length of crack required to give the observed fracture stress values is of the order of 10^{-5} cm. in length. This is large compared with the atomic spacing (10^{-8} cm.), so that for our calculations for the displacement of the first plane of atomic centres we can disregard powers of $(a/2c)$ above the lowest found in the expression for any stress or displacement. We then expect the approximate results so obtained to be correct to the order of one part in 10^3, provided that all the coefficients in the full expansions in terms of $(a/2c)^n$ are of an order not greater than that of the coefficient of the retained terms.

2.4. To find the displacement of the plane $z = \frac{1}{2}a$ we put $\zeta = a/2c = \zeta_0$ in the expressions of (2.1, 10). Thus we obtain to sufficiently high order in ζ_0:—

Domain 1. At great distances from the crack:

$$\left.\begin{aligned} &r \to \infty \quad \text{and so} \quad \zeta_0 \ll \rho^2 - 1, && S_0{}^0 = \tan^{-1}(\rho^2-1)^{-\frac{1}{2}},\\ &C_1{}^0 = (\rho^2-1)^{-\frac{1}{2}}, && S_1{}^0 = \zeta_0/(\rho^2-1)^{\frac{3}{2}},\\ &C_2{}^0 = \zeta_0{}^2(\rho^2-1)^{-\frac{3}{2}}, && S_0{}^1 = \rho^{-1}\{1-\zeta_0/(\rho^2-1)^{\frac{1}{2}}\},\\ &C_2{}^1 = \rho(\rho^2-1)^{-\frac{3}{2}}, && S_1{}^1 = \rho^{-1}(\rho^2-1)^{-\frac{1}{2}}. \end{aligned}\right\} \qquad \ldots\ldots(2.4,1)$$

Domain 2. Near the centre of the crack:

$$\left.\begin{aligned} &r \ll c \quad \text{and so} \quad \rho^2 \ll 1, && S_0{}^0 = \tan^{-1}[\zeta_0(1-\rho^2)^{\frac{1}{2}}],\\ &C_1{}^0 = \zeta_0(1-\rho^2)^{-\frac{3}{2}}, && S_1{}^0 = (1-\rho^2)^{-\frac{1}{2}},\\ &C_2{}^0 = -(1-\rho^2)^{-\frac{3}{2}}, && S_0{}^1 = \{1-(1-\rho^2)^{\frac{1}{2}} - \tfrac{1}{2}\zeta_0{}^2\rho^2(1-\rho^2)^{-\frac{3}{2}}\},\\ &C_2{}^1 = -3\zeta_0\rho(1-\rho^2)^{-\frac{3}{2}}, && S_1{}^1 = \zeta_0\rho(1-\rho^2)^{-\frac{3}{2}}. \end{aligned}\right\} \qquad \ldots\ldots(2.4,2)$$

Domain 3. Near the edge of the crack:

$r \simeq c$ and so $\rho^2 \simeq 1$; $\rho^2 - 1$ is of the same order as ζ_0.

Write
$$\rho - 1 = \mu\zeta_0. \qquad \ldots\ldots(2.4,3)$$

Then μ is of order unity in the domain where the following approximations hold. We may note that $\frac{1}{2}\mu$ is the number of lattice spacings between the point $(r, \frac{1}{2}a)$ or (ρ, ζ_0) being considered and the "edge" of the crack, $r = c$ $(\rho = 1)$, in the case of a simple cubic structure.

When

$$\left.\begin{aligned} &\rho - 1 = \mu\zeta_0,\\ &\quad C_1{}^0 = \zeta_0^{-\frac{1}{2}}\,2^{-1}(1+\mu^2)^{-\frac{1}{4}}\,[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}},\\ &\quad C_2{}^0 = \zeta_0^{-\frac{3}{2}}\,2^{-2}(1+\mu^2)^{-\frac{3}{4}}\,\{[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}} + \mu\,[1-\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}\},\\ &\quad C_2{}^1 = \zeta_0^{-\frac{3}{2}}\,2^{-1}(1+\mu^2)^{-\frac{3}{4}}\,\{\mu\,[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}} - [1-\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}\},\\ &\quad S_0{}^0 = \tan^{-1}\{2^{-1}(1+\mu^2)^{-\frac{1}{4}}\,[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{-\frac{1}{2}}\zeta_0^{-\frac{1}{2}}\},\\ &\quad S_1{}^0 = 2^{-1}(1+\mu^2)^{-\frac{1}{4}}\,\{1-\mu(1+\mu^2)^{-\frac{1}{2}}\}^{\frac{1}{2}}\zeta_0^{-\frac{1}{2}},\\ &\quad S_0{}^1 = 1-(1+\mu^2)^{\frac{1}{4}}\,\{1-\mu(1+\mu^2)^{-\frac{1}{2}}\}^{\frac{1}{2}}\zeta_0^{\frac{1}{2}}\rho,\\ &\quad S_1{}^1 = 2^{-1}(1+\mu^2)^{-\frac{1}{4}}\,\{1+\mu(1+\mu^2)^{-\frac{1}{2}}\}^{\frac{1}{2}}\zeta_0^{-\frac{1}{2}}\rho. \end{aligned}\right\} \qquad \ldots\ldots(2.4,4)$$

Putting these values in the expressions for u_z, σ_z, τ_{rz}, we have for $z=c\zeta_0$, $\zeta_0=a/2c$

$$u_z=\frac{T_0c\zeta_0}{E}-\frac{T_1\sigma c\zeta_0}{E}+\frac{4(1-\sigma^2)cT_0}{\pi E}U_z, \qquad \ldots\ldots(2.4,5)$$

where
$$U_z=(1-\rho^2)^{\frac{1}{2}}-\frac{1-2\sigma}{2(1-\sigma)}\zeta_0\tan^{-1}[\zeta_0^{-1}(1-\rho^2)^{\frac{1}{2}}] \qquad \ldots\ldots(2.4,6)$$

if $\rho\ll 1$, i.e. near centre of crack,

$$U_z=\left\{[1-\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}-\frac{1}{4(1-\sigma)}[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}(1+\mu^2)^{-\frac{1}{2}}\right\}\zeta^{\frac{1}{2}} \qquad \ldots\ldots(2.4,7)$$

if $\rho-1=\mu\zeta_0$, $\mu=0(1)$, i.e. at edge of crack,

$$U_z=\frac{1-2\sigma}{2(1-\sigma)}\left\{\frac{1}{(\rho^2-1)^{\frac{1}{2}}}-\tan^{-1}\frac{1}{(\rho^2-1)^{\frac{1}{2}}}\right\}\zeta_0 \qquad \ldots\ldots(2.4,8)$$

if $\rho\gg 1$ at great distances from the crack.

$$\sigma_z=T_0+\frac{2T_0}{\pi}\Sigma_z, \qquad \ldots\ldots(2.4,9)$$

where
$$\Sigma_z=\tan^{-1}[\zeta_0^{-1}(1-\rho^2)^{\frac{1}{2}}], \qquad \ldots\ldots(2.4,10)$$

if $\rho\ll 1$,

$$\Sigma_z=(1+\mu^2)^{-\frac{3}{2}}\{[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}[1+2(1+\mu^2)]+\mu[1-\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}\}\zeta_0^{-\frac{1}{2}}/4 \qquad \ldots\ldots(2.4,11)$$

if $\rho-1=\mu\zeta_0$, $\mu=0(1)$,

$$\Sigma_z=(\rho^2-1)^{-\frac{1}{2}}-\sin^{-1}(1/\rho) \qquad \ldots\ldots(2.4,12)$$

if $\rho\gg 1$, and
$$\tau_{rz}=\frac{2T_0}{\pi}T_{rz}, \qquad \ldots\ldots(2.4,13)$$

where
$$T_{rz}=-\zeta_0{}^2\rho(1-\rho^2)^{-\frac{3}{2}}\{1+3(1-\rho^2)^{-1}\}, \qquad \ldots\ldots(2.4,14)$$

if $\rho^2\ll 1$,

$$T_{rz}=(1+\mu^2)^{-\frac{3}{2}}\{\mu[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}-[1-\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}\}\zeta_0^{-\frac{1}{2}}/4, \qquad \ldots\ldots(2.4,15)$$

if $\rho-1=\mu\zeta_0$,

$$T_{rz}=\zeta_0\{\rho(\rho^2-1)^{-\frac{3}{2}}-\rho^{-1}(\rho^2-1)^{-\frac{1}{2}}\} \qquad \ldots\ldots(2.4,16)$$

if $\rho^2\gg 1$.

2.5. We are now able to compute the displacement u_z and the stresses σ_z and τ_{rz} due to any T_0, length of crack, c and atomic spacing a, i.e. any ζ_0. We are particularly interested in the region near the edge of the crack ($\rho\simeq 1$) (*Domain* 3), i.e. in the domain $-10\leqslant\mu\leqslant+10$ and especially $-1\leqslant\mu\leqslant+1$. The results for this domain are therefore given in detail (table 1, where values of A, C, D and F are given), u_z, σ_z and τ_{rz} being given by

$$u_z=\frac{T_0c\zeta_0}{E}-\frac{\sigma T_1c\zeta_0}{E}+\frac{4(1-\sigma^2)aT_0}{2\pi E}\left\{C-\frac{D}{4(1-\sigma)}\right\}\zeta_0^{-\frac{1}{2}}, \qquad \ldots\ldots(2.5,1)$$

$$\sigma_z=T_0+\frac{2T_0}{\pi}A\zeta_0^{-\frac{1}{2}}, \qquad \ldots\ldots(2.5,2)$$

$$\tau_{rz}=\frac{2T_0}{\pi}F\zeta_0^{-\frac{1}{2}}, \qquad \ldots\ldots(2.5,3)$$

for $\rho-1=\mu\zeta_0$, $-10\leqslant\mu\leqslant+10$.

Table 1

μ	r	A	C	D	F
−10	$c-5{\cdot}0\,a$	+0·00004	+4·4775	+0·02204	−0·00165
− 9	$c-4{\cdot}5\,a$	0·00026	4·2493	0·02602	−0·00214
− 8	$c-4{\cdot}0\,a$	0·00033	4·0076	0·03100	−0·00255
− 7	$c-3{\cdot}5\,a$	0·00048	3·7505	0·03771	−0·00400
− 6	$c-3{\cdot}0\,a$	0·00081	3·4761	0·04731	−0·00578
− 5	$c-2{\cdot}5\,a$	0·00105	3·1774	0·06171	−0·00895
− 4	$c-2{\cdot}0\,a$	0·00314	2·8495	0·08489	−0·01513
− 3	$c-1{\cdot}5\,a$	0·00802	2·4825	0·12742	−0·02918
− 2	$c-1{\cdot}0\,a$	0·02078	2·0593	0·21726	−0·06775
− 1·0	$c-0{\cdot}5\,a$	0·14716	1·5536	0·45516	−0·19422
− 0·4	$c-0{\cdot}2\,a$	0·44936	1·21556	0·76386	−0·30907
− 0·2	$c-0{\cdot}1\,a$	0·60528	1·10460	0·88784	−0·30305
− 0·1	$c-0{\cdot}05\,a$	0·68016	1·05230	0·94536	−0·28259
+ 0·0	$c-0{\cdot}00\,a$	0·75000	1·00000	1·00000	−0·25000
+ 0·1	$c+0{\cdot}05\,a$	0·80612	0·95006	1·04708	−0·20808
+ 0·2	$c+0{\cdot}1\,a$	0·84560	0·90544	1·08064	−0·16135
+ 0·4	$c+0{\cdot}2\,a$	0·87329	0·82288	1·12832	−0·06236
+ 0·6	$c+0{\cdot}3\,a$	0·85027	0·75248	1·14216	+0·02770
+ 1·0	$c+0{\cdot}5\,a$	0·74354	0·64370	1·09850	+0·07541
+ 2	$c+1{\cdot}0\,a$	0·51600	0·48582	0·92053	+0·08119
+ 3	$c+1{\cdot}5\,a$	0·42165	0·40297	0·78501	+0·05569
+ 4	$c+2{\cdot}0\,a$	0·36080	0·35003	0·69109	+0·03949
+ 5	$c+2{\cdot}5\,a$	0·31327	0·31468	0·62312	+0·02932
+ 6	$c+3{\cdot}0\,a$	0·29150	0·28770	0·57147	+0·02285
+ 7	$c+3{\cdot}5\,a$	0·26909	0·26667	0·52916	+0·01837
+ 8	$c+4{\cdot}0\,a$	0·25142	0·24996	0·49706	+0·01517
+ 9	$c+4{\cdot}5\,a$	0·23672	0·23558	0·46925	+0·01280
+10	$c+5{\cdot}0\,a$	0·22444	0·22150	0·44556	+0·01098

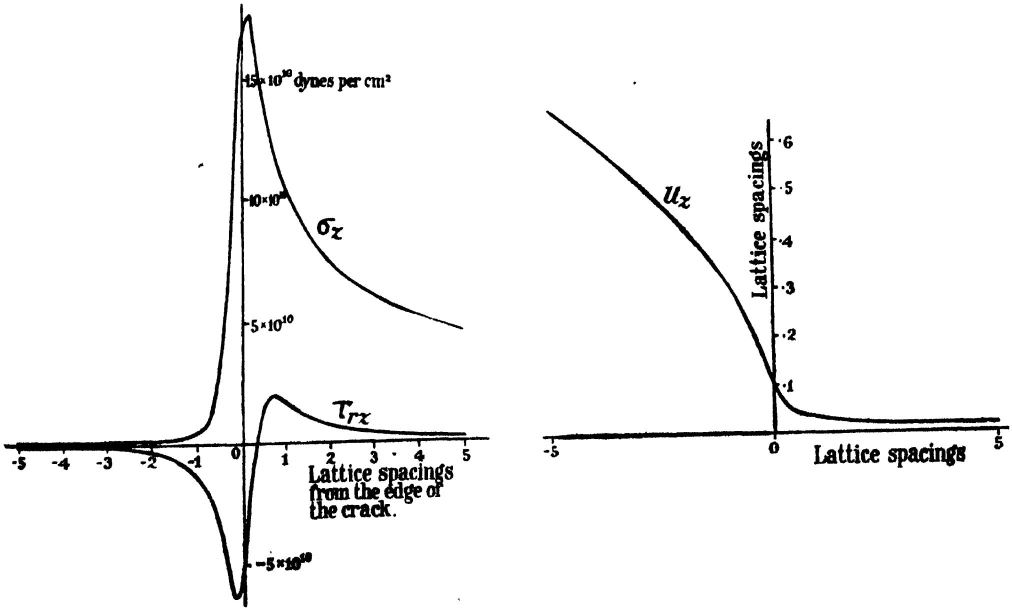


Figure 3. Normal stress, shear stress and normal displacement for planes $z = \pm a$, $T_0 = 10^{10}$ dynes/cm.^{2}, $E = 10^{12}$ dynes/cm.^{2}, $\sigma = 0{\cdot}25$. $2c = 1000a$.

The displacement u_z and distribution of the stresses along the first plane of atomic centres (for $T_1=0$, $T_0=10^{10}$ dynes/sq. cm., $E=10^{12}$ dynes/cm. $\sigma=0{\cdot}25$) is shown in figure 3 where σ_z, τ_{rz} and u_z are shown for distances measured in terms of atomic radii from the edge of the crack, for a crack 1000 lattice spacings in length (see also table 2).

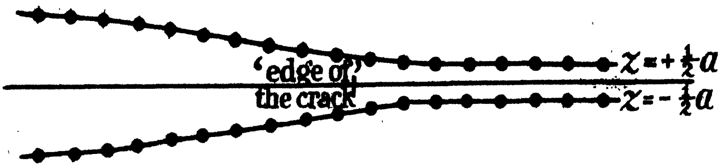


Figure 4. Configuration of atoms near the edge of the crack.

In figure 4 the displacement is shown for a crack about 2500 lattice spacings long in a glass typical of that used by Griffith (1921); this length is approximately that required to give the rupture strength observed by Griffith.

It will be seen that the change-over between the three regions is smooth and occurs about 5 lattice spacings from the crack edge. This is shown in table 2 for a crack 1000 atoms long, the values of U_z, Σ_z and T_{rz} being given as calculated from the expressions valid in each of the three regions.

Table 2

$\rho=\frac{r}{c}$	μ	U_z			Σ_z			T_{rz}		
		$\rho\ll1$	$\rho\simeq1$	$\rho\gg1$	$\rho\ll1$	$\rho\simeq1$	$\rho\gg1$	$\rho\ll1$	$\rho\simeq1$	$\rho\gg1$
0·800	−200	0·5995			−1·569			−0 ·000		
0·850	−150	0·5263			−1·569			− 0·000		
0·900	−100	0·4354			−1·569			− 0·000		
0·950	− 50	0·3118			−1·568			− 0·001		
0·990	− 10	0·1406	0·1408		−1·564	− 1·563		− 0·054	− 0·052	
0·992	− 8	0·1258	0·1259		−1·563	− 1·553		− 0·095	− 0·090	
0·994	− 6	0·1089	0·1089		−1·562	− 1·536		− 0·191	−0·183	
0·996	− 4	0·0888	0·0887		−1·559	− 1·461		− 0·527	−0·480	
0·997	− 3	0·0769	0·0766		−1·558	− 1·304		1·079	−0·922	
0·998	− 2	0·0628	0·0620		−1·555	− 0·898		− 2·962	− 2·147	
0·999	− 1	0·0442	0·0438		−1·548	+ 4·103		−16·772	− 6·139	
1·000	0		0·0206			+22·179			−7·902	
1·001	+ 1		0·0083	0·0068		+21·980	+20·29		+ 2·384	+10·377
1·002	+ 2		0·0052	0·0047		+14·812	+13·92		+2·566	+ 3·666
1·003	+ 3		0·0040	0·0037		+11·842	+11·11		+ 1·760	+ 1·993
1·004	+ 4		0·0033	0·0031		+ 9·928	+ 9·43		+1·249	+ 1·293
1·006	+ 6		0·0025	0·0025		+ 7·757	+ 7·44		+0·722	+ 0·701
1·008	+ 8		0·0021	0·0021		+ 6·505	+ 6·274		+0·479	+ 0·455
1·010	+ 10		0·0018	0·0018		+ 5·667	+ 5·628		+0·347	+ 0·348
1·050	+ 50			0·0006			+ 1·863			+ 0·029
1·100	+100			0·0004			+ 1·041			+ 0·009
1·150	+150			0·0002			+ 0·707			+ 0·004
1·200	+200			0·0002			+ 0·522			+ 0·003

§3. ANALYSIS OF STRESS AND STRAIN FOR THE TWO-DIMENSIONAL CRACK

3.1. The problem of the crack formed by a cut from $x=-c$ to $x=+c$ in an elastic medium is easily seen to be equivalent to the two-dimensional problem of an infinite semi-plane with the boundary conditions

$$\left.\begin{array}{lll}\tau_{xz}=0 & \text{on} & z=0, \\ \sigma_z=-p_0 & \text{on} & z=0,\ |x|<c, \\ u_z=0 & \text{on} & z=0,\ |x|\geqslant c,\end{array}\right\} \qquad \ldots\ldots(3.1,1)$$

if the crack is maintained by an internal pressure.

We use the solution of this problem as found by Westergaard (1939) and Elliott and Sneddon (1947). Changing the problem to that of a crack in a body in tension (T_0) as we did in the three-dimensional case, we have the solutions

$$\tfrac{1}{2}(\sigma_x+\sigma_z)=T_0\left\{\frac{r}{r_1^{\frac{1}{2}}r_2^{\frac{1}{2}}}\cos\left(\theta-\tfrac{1}{2}\theta_1-\tfrac{1}{2}\theta_2\right)\right\}, \qquad \ldots\ldots(3.1,2)$$

$$\tfrac{1}{2}(\sigma_z-\sigma_x)=T_0\frac{r\sin\theta}{c}\left(\frac{c^2}{r_1r_2}\right)^{\frac{3}{2}}\sin\tfrac{3}{2}(\theta_1+\theta_2), \qquad \ldots\ldots(3.1,3)$$

$$\tau_{xz}=T_0\frac{r\sin\theta}{c}\left(\frac{c^2}{r_1r_2}\right)^{\frac{3}{2}}\cos\tfrac{3}{2}(\theta_1+\theta_2), \qquad \ldots\ldots(3.1,4)$$

where $\quad x+iz=re^{i\theta},\quad x-c+iz=r_1e^{i\theta_1},\quad x+c+iz=r_2e^{i\theta_2}, \qquad \ldots\ldots(3.1,5)$

and for the "shape" of the crack we have the ellipse

$$\frac{x^2}{c^2}+\left(\frac{u_z}{\epsilon}\right)^2=1, \qquad \ldots\ldots(3.1,6)$$

where $\quad \epsilon=2(1-\sigma^2)T_0c/E. \qquad \ldots\ldots(3.1,7)$

By comparison with (2.3,2) we see that u_z for the two-dimensional problem will be $\pi/2$ times that for the three-dimensional case with the same applied tensions.

3.2. We now proceed to find the stresses and displacements of the first planes of atomic centres from the cleavage plane. As before, we write $\zeta_0=a/2c$ and treat ζ_0 as small.

Let us first find the stresses in the neighbourhood of the edge of the crack.

If μ is measured from the crack edge as in the three-dimensional case, i.e. $(x/c)-1=\mu\zeta_0$, we have

$$\left.\begin{array}{r}r_2\simeq 2c+\mu a/2, \\ r\simeq c+\mu a/2, \\ r_1\cos\theta_1\simeq x_1=\mu a, \\ r_2\cos\theta_2\simeq z_1=a/2,\end{array}\right\} \qquad \ldots\ldots(3.2,1)$$

and so

$$\tfrac{1}{2}(\sigma_x+\sigma_z)=T_0\zeta_0^{-\frac{1}{2}}\frac{[(1+\mu^2)^{\frac{1}{2}}+\mu]^{\frac{1}{2}}}{2(1+\mu^2)^{\frac{1}{2}}}, \qquad \ldots\ldots(3.2,2)$$

$$\tfrac{1}{2}(\sigma_z-\sigma_x)=T_0\zeta_0^{-\frac{1}{2}}\frac{[(1+\mu^2)^{\frac{1}{2}}+\mu]^{\frac{1}{2}}+\mu[(1+\mu^2)^{\frac{1}{2}}-\mu]^{\frac{1}{2}}}{4(1+\mu^2)^{\frac{3}{2}}}, \qquad \ldots\ldots(3.2,3)$$

giving
$$\sigma_z = T_0\zeta_0^{-\frac{1}{2}}\frac{[(1+\mu^2)^{\frac{1}{2}}+\mu]^{\frac{1}{2}}[1+2(1+\mu^2)]+\mu[(1+\mu^2)^{\frac{1}{2}}-\mu]^{\frac{1}{2}}}{4(1+\mu^2)^{\frac{3}{2}}}, \qquad \ldots\ldots(3.2,4)$$

$$\sigma_x = T_0\zeta_0^{-\frac{1}{2}}\frac{[(1+\mu^2)^{\frac{1}{2}}+\mu]^{\frac{1}{2}}[2(1+\mu^2)-1]-\mu[(1+\mu^2)^{\frac{1}{2}}-\mu]^{\frac{1}{2}}}{4(1+\mu^2)^{\frac{3}{2}}}, \qquad \ldots\ldots(3.2,5)$$

and
$$\tau_{xz} = T_0\zeta_0^{-\frac{1}{2}}\frac{\mu[(1+\mu^2)^{\frac{1}{2}}+\mu]^{\frac{1}{2}}-[(1+\mu^2)^{\frac{1}{2}}-\mu]^{\frac{1}{2}}}{4(1+\mu^2)^{\frac{3}{2}}}. \qquad \ldots\ldots(3.2,6)$$

By comparison with the results for the three-dimensional case we see that

$$\left.\begin{aligned}[\sigma_z]_3 &= \frac{2}{\pi}[\sigma_z]_2,\\ [\tau_{rz}]_3 &= \frac{2}{\pi}[\tau_{rz}]_2,\end{aligned}\right\} \qquad \ldots\ldots(3.2,7)$$

in this region, and as we have seen above (3.1),

$$[u_z]_3 = \frac{2}{\pi}[u_z]_2.$$

To this accuracy then

$$[\sigma_z]_2 = T_0A\zeta_0^{-\frac{1}{2}}, \qquad \ldots\ldots(3.2,8)$$

$$[\tau_{rz}]_2 = T_0F\zeta_0^{-\frac{1}{2}}, \qquad \ldots\ldots(3.2,9)$$

$$[u_z]_2 = \frac{(1-\sigma^2)aT_0}{E}\left\{C-\frac{D}{4(1-\sigma)}\right\}\zeta_0^{-\frac{1}{2}}, \qquad \ldots\ldots(3.2,10)$$

where A, C, D, F have the same values as in the three-dimensional case. Hence from any results which we derive for the three-dimensional case with approximations of this accuracy, we can obtain the results for the two-dimensional case by using the multiplying factor $\frac{1}{2}\pi$.

§4. CONSTRUCTION OF A MODEL AND THE DERIVATION OF CONDITIONS FOR RUPTURE

4.1. We now use the results of the above analysis to find a suitable model for a crack in a real solid.

We may replace our infinite ideal elastic solid under tension with the theoretical crack $x^2+y^2 \leqslant c^2$, $z=0$ (or $|x| \leqslant c$, $z=0$) by two semi-infinite blocks originally bounded by the planes $z=\pm\frac{1}{2}a$ and a law of force between the two planes such that it maintains forces equal to $[\sigma_z]_{z=\pm\frac{1}{2}a}$ and $[\tau_{rz}]_{z=\pm\frac{1}{2}a}$ on the two planes. As τ_{rz} is small compared with σ_z and is along the surface, we shall neglect it; we shall suppose that, if σ_z is plotted against μ_z, and the resulting function is taken as the law of force, the strains are approximately those calculated above.

If we graph σ_z against u_z we find that for $c \simeq 10^4a$ the form of the curve between $u_z = 0{\cdot}05a$ and a is given by

$$\left.\begin{aligned}\sigma_z &= \frac{2T_0}{\pi}A\zeta_0^{-\frac{1}{2}},\\ u_z &= \frac{4(1-\sigma^2)}{\pi E}\cdot\frac{a}{2}\cdot T_0\left\{C-\frac{D}{4(1-\sigma)}\right\}\zeta_0^{-\frac{1}{2}}\end{aligned}\right\} \qquad \ldots\ldots(4.1,1)$$

in the case of the three-dimensional crack (see figure 5), and

$$\left.\begin{aligned} \sigma_z &= T_0 A \zeta_0^{-\frac{1}{2}}, \\ u_z &= \frac{2(1-\sigma^2)}{E} \cdot \frac{a}{2} \cdot T_0 \left\{ C - \frac{D}{4(1-\sigma)} \right\} \zeta_0^{-\frac{1}{2}} \end{aligned}\right\} \quad \ldots\ldots(4.1, 2)$$

in the two-dimensional case, i.e. by the expressions which hold near the edge of the crack, to an accuracy better than 1%.

It will be noted that the above expressions depend only on a, E, σ and $T_0\sqrt{(2c/a)}$ as parameters, A, C and D being independent of any properties of the body.

If we consider σ_z for very small values of u_z we find that $\sigma_z/(2u_z/a)$ tends to the value E. This is to be expected, as for large distances from the crack, i.e. very small u_z, the extension is merely given by Hooke's law for simple tension. For large values of u_z, $\sigma_z \sim \left[1 - \frac{2}{\pi}\tan^{-1}\lambda u_z\right]$ and tends to zero as u_z tends to infinity.

We thus have a law of force, σ_z as a function of z, which behaves in the manner of the interatomic forces; the initial rate of increase of stress with strain is E, there exists a maximum attractive force, and the stress dies away to zero for infinite strain. If, in fact, the law of force we have derived is to correspond to the true interatomic forces expected, then the maximum stresses (P) must be equal. This condition determines the parameter $T_0\sqrt{(2c/a)}$, i.e. $T_0\zeta_0^{-\frac{1}{2}}$ appearing in (4.1, 1) or (4.1, 2), in fact

$$T_0\sqrt{(2c/a)} = \frac{\pi}{2A_{\max}} P \quad \text{from the three-dimensional case,} \quad \ldots\ldots(4.1, 3)$$

or

$$T_0\sqrt{(2c/a)} = \frac{1}{A_{\max}} P \quad \text{from the two-dimensional case,} \quad \ldots\ldots(4.1, 4)$$

and (4.1, 1) and (4.1, 2) are replaced by

$$\left.\begin{aligned} \sigma_z &= P \,.\, A/A_{\max}, \\ u_z &= \frac{2(1-\sigma^2)a}{2} \cdot \frac{P}{E} \cdot \left\{ C - \frac{D}{4(1-\sigma)} \right\} / A_{\max} \end{aligned}\right\} \quad \ldots\ldots(4.1, 5)$$

for either crack. It will be noted that if we do not impose this condition on $T_0\sqrt{(2c/a)}$ we cannot regard our solution, based on the pure elastic theory, as being in any agreement with the true solution at the ends of a crack based on interatomic or intermolecular considerations. For if T_0 or c is increased above the values for (4.1, 3) or (4.1, 4) to hold, then the maximum of our stress-strain curve is greater than the maximum theoretical stress P and so we cannot have a state of equilibrium and rupture must have commenced. Similarly, if T_0 or c is reduced, we cannot have equilibrium and the crack must be closing to some other shape. In fact, such a crack would close up completely unless some inclusion existed to maintain it.

Also it will be observed that conditions (4.1, 3) and (4.1, 4) are the conditions for the corresponding cracks to produce rupture. The conditions of (4.1, 3) and (4.1, 4) are, in fact, the equivalents of the usual criteria given by Griffith (1924).

Inserting the numerical values in (4.1, 3) and (4.1, 4), we have:—

For rupture,

$$\left.\begin{aligned} T_0\sqrt{c} &= 1{\cdot}269\,P\sqrt{a} \quad \text{for penny-shaped crack,} \\ T_0\sqrt{c} &= 0{\cdot}808\,P\sqrt{a} \quad \text{for a “Griffith” crack.} \end{aligned}\right\} \quad \ldots\ldots(4.1, 6)$$

The form of the law of force derived is shown in figure 5; it is plotted in the most suitable non-dimensional form, σ_z/P against $\frac{E}{P}\cdot\frac{2u_z}{a}$, σ being put equal to 0·25.

We have thus far identified our law of stress-strain with a law expected from interatomic attractive forces only by the equivalence of the maxima, the slope at zero strain, and a diminution of the stress to zero for infinite strain. In fact we have determined the stress-strain law (assuming P is known) by this equivalence, as we had only the one parameter $T_0\sqrt{2c/a}$ in our expression for stress and strain. This makes us unable to obtain exact agreement between the stress-strain law derived as above and that expected from interatomic forces.

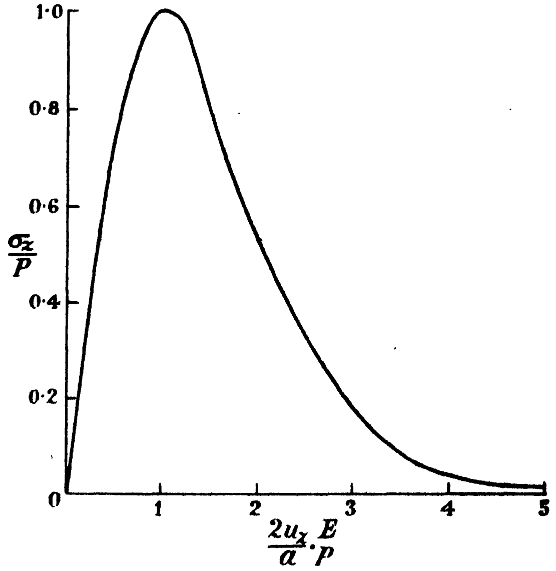


Figure 5. σ_z as a function of u_z.

The graph is plotted in the dimensionless form $\frac{\sigma_z}{P}$ against $\frac{2u_z}{a}\cdot\frac{E}{P}$.

There are two main differences:—Firstly, in the stress-strain law derived above, the slope increases with strain from E to $1\cdot5E$ (approx.) before decreasing to zero at the stress maximum; this in turn causes the stress maximum to be attained for a smaller strain than that expected from interatomic laws of force. Secondly, the stress should tend to zero for large strain as $K/(2u_z+a)^3$, i.e., a van der Waals law of force (De Boer, 1936), while the stress-strain relationship found above gives $\tan^{-1}(\lambda u_z)$ which is a logarithmic decay. However, we know from the ideal elastic solution that σ_z would be zero on a surface only $\frac{1}{2}a$ from the first plane of atom centres, so that for large u_z (near centre of the crack) our model will not be greatly in error.

4.2. One other result can be derived from our law of force.

Our assumed law of force allows us to find the surface energy of the body. It is simply

$$\left.\begin{aligned} S &= \int_0^{\infty} \sigma_z d(u_z), \\ &= \int_0^{a/2} \sigma_z d(u_z) + \int_{a/2}^{\infty} \sigma_z d(u_z), \end{aligned}\right\} \qquad \ldots\ldots(4.2, 1)$$

$$= I_1 + I_2. \qquad \ldots\ldots(4.2, 2)$$

I_1 can be calculated numerically in each case using table 1, or from figure 5, once the values of P, E, σ and a are known. I_2 can be calculated, as a small correction, by using the fact that in this region σ_z is due to van der Waals's forces which die away as $K/(a+2u_z)^3$, K being determined by the value of σ_z at $u_z = \frac{1}{2}a$; I_2 is a correction of the order of 1% or less:

$$I_2 = a[\sigma_z]_{u_z = \frac{1}{2}a}.$$

Alternatively the integral may be found to sufficient accuracy by transforming it to the integral

$$S = \frac{(1-\sigma^2)P^2 a}{E(A_{\max})^2} \int_0^a A d\left\{C - \frac{D}{4(1-\sigma)}\right\}, \qquad \ldots\ldots(4.2, 3)$$

which from figure 5 gives ($\sigma = 0{\cdot}25$)

$$S = 0{\cdot}92 \frac{P^2 a}{E}. \qquad \ldots\ldots(4.2, 4)$$

This enables us to put our criterion for rupture in a form analogous to that of Griffith (1921),

$$T_0\sqrt{c} = 0{\cdot}84\sqrt{(S \,.\, E)}. \qquad \ldots\ldots(4.2, 5)$$

4.3. Unfortunately the only results which allow full comparison of theory with experiment are those of Griffith, which are for glass—a non-crystalline substance. If, however, we assume that the crack is such that at least at its ends it is equivalent to a cut in a silica crystal along a basal plane and the mechanism of fracture is to cause cleavage to spread along this plane, we can use the above analysis. In fact, we assume that fracture begins with the spread of the edge of the crack through a crystallite of silica.

We then have

$E = 6{\cdot}34 \times 10^{11}$ dynes/sq. cm. from Griffith,

$S_{\text{expt}} = 546$ dynes/cm. from Griffith (by extrapolation from the liquid phase),

and the normal rupture strength for his glass

$$= 1{\cdot}83 \times 10^9 \text{ dynes/sq. cm.}$$

For quartz we have $a = 5 \times 10^{-8}$ cm. $= 1{\cdot}97 \times 10^{-8}$ in.

Griffith (1921) performed a series of experiments to determine $T_0\sqrt{c}$ for cracks produced artificially with a diamond. His values varied with annealing temperature, etc. (see Griffith, 1924), but by careful considerations of the factors involved, he finally found consistent values for $T_0\sqrt{c}$ of $1{\cdot}43 \times 10^7$, compared with $2{\cdot}69 \times 10^7$ of his earlier paper (T_0 in dynes/cm^2, c in cm.).

This gives us a value of $P = 8 \times 10^{10}$ dynes/sq. cm., a result which is slightly lower than that of $1{\cdot}3 \times 10^{11}$ given by Griffith (1924) by taking the radius of curvature of the end of the crack as 5×10^{-8} cm.

Using this value of $P = 8 \times 10^{10}$ dynes/sq. cm. we find that for the normal rupture strength the glass must contain cracks of half-length $c = 6{\cdot}1 \times 10^{-5}$ cm.

We may also use this value of P to find S, the surface tension. We find

$$S = 466 \text{ dynes/cm.},$$

which is somewhat lower than the value given by Griffith (1924).

§ 5. POSSIBLE PHYSICAL FORMS OF A CRACK

5.1. We may now examine our model to see whether we can avoid the mathematical fiction that when tensions are removed the crack is merely a plane across which no cohesive forces can act. For the above analysis to hold, it is sufficient that the unstrained state of the body should be that of a perfect solid lattice, and that for tensions sufficiently large for rupture, the crack surfaces should be free from forces (i.e. hydrostatic pressures, etc.). Our analysis does not forbid an inclusion that does not fill the space within the crack at rupture, and does not itself adhere strongly to the surfaces of the crack, i.e. a stress less than the rupture stress will separate it from the surface of the crack. This would obviously permit gaseous inclusions. Examination of the shape of the crack at rupture shows that the included atoms could at least extend to within two or three atom diameters of the "edge", and yet allow the surface to be "free" at rupture. If the included plane of atoms is such that the inclusion is solid or liquid, and does not adhere strongly, the only surface forces at rupture on the crack will be slight vapour pressures. If the included atoms are gaseous, the pressures may be greater.

5.2. An alternative hypothesis is to assume that one plane of atoms is missing between the faces of the crack, the penny-shaped hole being filled by a gas or other inclusion. Such an inclusion is necessary to maintain the hole; if the inclusion is not present, rough analyses show that cracks greater than about 10–20 atoms long close spontaneously or under thermal agitations. The hypothesis of such a crack does not greatly increase the size of crack required for rupture. This may be seen as follows :—

The stresses on the first plane of atomic centres in this case will be less than those given above as the plane is now distant a from the plane of symmetry.

Let us assume that to a first approximation it is equal to the stress calculated by the former analysis for the planes $z = \pm a$. Then

$$\left.\begin{aligned}(\sigma_z)_{z=a} &= \frac{2T_0}{\pi} A \left(\frac{a}{c}\right)^{\frac{1}{2}},\\ &\simeq \frac{1}{\sqrt{2}} \cdot \frac{2T_0}{\pi} A \zeta_0^{-\frac{1}{2}}.\end{aligned}\right\} \qquad \ldots\ldots(5.1)$$

Also we have no longer to equate the maximum value of this to P, the maximum cohesive stress which is theoretically possible for a perfect solid, for the atomic plane at distance a is missing in the crack itself. If we estimate the effect of this, using a suitable law of interatomic force, we find that if P_1 is the new maximum attractive force, then

$$P_1 \simeq 0{\cdot}8P. \qquad \ldots\ldots(5.2)$$

Hence we have for rupture

$$\left.\begin{aligned} &P \simeq T_0 \frac{2A_{\max}}{\pi} \zeta_0^{-\frac{1}{2}} \cdot \frac{0{\cdot}70}{0{\cdot}80}, \\ \text{or} \quad &P = 0{\cdot}487\, T_0 \zeta_0^{-\frac{1}{2}}. \end{aligned}\right\} \qquad \ldots\ldots(5.3)$$

For the glass used by Griffith, a crack about 3600 atoms long is now required to account for observed rupture stress.

As before, a "non-adhering" solid or liquid inclusion is permissible and has no appreciable effect on the analysis given above for rupture. We may also have a gaseous inclusion if it is not strongly absorbed by the material.

It will be seen that the effect on the rupture strength will be the same even if the inclusion has a less regular shape than that suggested above, i.e. a single plane of atoms, provided that it is only a single plane near the edges of the crack. "Near" in this case may be of the order of only 10–100 atom diameters, provided the inclusion does not broaden too rapidly beyond this distance.

ACKNOWLEDGMENTS

The author wishes to express his thanks to Professor N. F. Mott, F.R.S. and Mr. F. R. N. Nabarro for several illuminating and helpful discussions during the preparation of this paper, and to Imperial Chemical Industries for financial support.

REFERENCES

de Boer, J. H., 1936. *Trans. Faraday Soc.*, **32**, 10.
Elliott, H. A. and Sneddon, I. N., 1947. (In the press.)
Griffith, A. A., 1921. *Phil. Trans. Roy. Soc.*, A, **222**, 180.
Griffith, A. A., 1924. *Proc. Int. Congr. Appl. Mech.* (Delft), 55.
Inglis, C. E., 1911. *Trans. Inst. Nav. Arch.*, **55**, 219.
Sneddon, I. N., 1947. *Proc. Roy. Soc.* (In the press.)
Westergaard, H. M., 1939. *J. Appl. Mech.*, **6**, 49.

MULTIPLE-BEAM LOCALIZED FRINGES: PART I.—INTENSITY DISTRIBUTION AND LOCALIZATION

By J. BROSSEL,
Paris
(Now at Manchester University)

Communicated by S. Tolansky; MS. received 4 *September* 1946

ABSTRACT. The interference taking place when a parallel beam of light is incident on a wedge with highly reflecting surfaces is studied theoretically, an approximation correct to the third order being used and non-normal incidence considered. In addition to the well known plane of fringe localization close to the wedge, there is an infinity of other planes where clearly resolved fringes are observed even at distances up to several metres from the interferometer. In most cases the spacing of the maximum is a sub-multiple of that given by the classical formula $N\lambda = 2t \cos \theta$, but if minor details are taken into account the fringe period is the same in all planes of localization. This phenomenon is repetitive at equal intervals along the optical axis. All the effects have been experimentally observed. The significance of two fringe-broadening terms in the analysis is explained with reference to practical cases and the importance of small gaps is stressed. The theory predicts that a similar repetitive law of non-classical fringe spacing should exist at large distances along the line of greatest slope of the wedge, i.e. at large gaps of a centimetre or more.

§ 1. INTRODUCTION

Two types of low-order multiple-beam interference fringes have recently been extensively applied to the study of surface topography (Tolansky, 1943 a):

(*a*) "Fizeau" (localized) fringes of equal thickness formed by monochromatic light between two surfaces of reflection coefficient 85% to 95%.

(*b*) Fringes of equal chromatic order formed in a similar interferometer with white light, being channelled spectra with Fabry-Perot type intensity distribution (Tolansky, 1945).

The methods have ready application, for the fringes are easily obtained with high intensity and extreme sharpness.

It is the purpose of this paper to discuss the formation of these fringes and draw attention to novel properties predicted by theory and confirmed by experiment.

§ 2. PATH DIFFERENCE AFTER MULTIPLE REFLECTION

Figure 1 shows the arrangement used for Fizeau fringes. A monochromatic point source S is situated at the focus of a lens C and the resulting parallel beam falls on the interferometer I, which consists of two highly reflecting slightly transparent surfaces close together, on which fringes of equal thickness are formed. An objective M projects the fringes on to the plane E. If white light is used and a spectrograph slit brought into plane E, fringes of equal chromatic order appear in the spectral plane.

The case of a simple wedge is easily analysed mathematically, as shown below. For the corresponding experimental observations, two silvered Fabry-Perot flats were used with thin mica spacers.

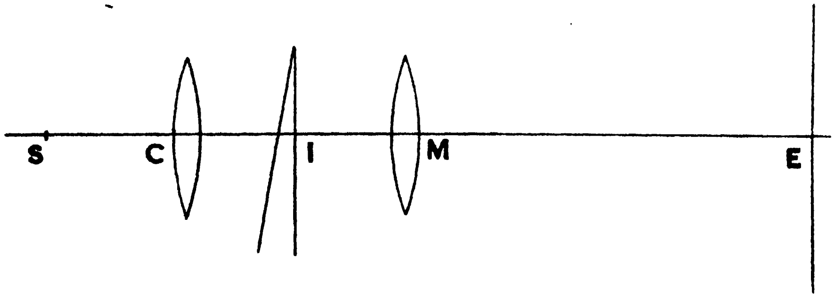


Figure 1.

Figure 2 represents a wedge, of which one reflecting plane is OY and the other is at a small angle ϵ to this; their line of intersection passes through O perpendicular to the plane of the figure. Let the intensity reflection coefficient of each plane be r. A plane wave approaches from the left at an angle of incidence θ and suffers multiple reflection inside the interferometer. This gives rise to a family of plane waves $\pi_0, \pi_1 \ldots . \pi_n \ldots .$, the combination of which defines the state of interference at the point P (X_0, Y_0). These plane waves are in phase, have intensities decreasing geometrically, and the angle between successive members is 2ϵ.

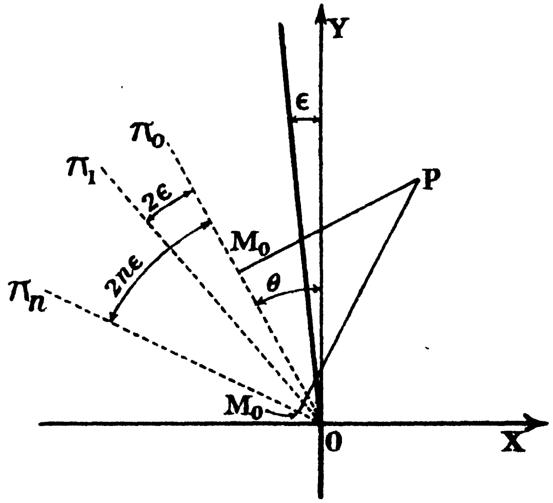


Figure 2.

The resultant vibration at P (X_0, Y_0) is given by

$$A \cos \Delta = \sum_n a r^n \cos \phi_n. \qquad \ldots\ldots(1)$$

This expression depends only on the position of P relative to the family $\pi_0, \pi_1 \ldots . \pi_n \ldots .$ The resultant at P is the same for all positions of the interferometer which do not affect the multiply-reflected wave-fronts π_0, π_1.... π_n...., i.e. states produced by the rotation of the wedge about its ridge. The path difference between the direct beam arriving at P and the beam which has suffered $2n$ reflections is

$$\delta_n = \mathrm{PM}_n - \mathrm{PM}_0 = X_0[\cos(\theta+2n\epsilon)-\cos\theta] + Y_0[\sin(\theta+2n\epsilon)-\sin\theta]. \qquad ......(2)$$

The terms of the series to be summed to give the resultant vibration A decrease rapidly, and we now consider the number of terms to be taken into account to obtain a given approximation.

Whatever the phase law, the series (1) is absolutely and uniformly convergent and its sum is always less than $a/(1-r)$. If only the first n terms are taken into account, the error R_n in the amplitude is less than $ar^n/(1-r)$ and the relative error in the fringe intensity distribution is

$$\frac{\Delta I}{I} = 2\left|\frac{R_n}{A}\right| \leqslant \frac{2ar^n}{(1-r)|A|}.$$

Hence if we wish to calculate the intensity with a relative accuracy better than one per cent, above the ordinate $I_0 = A_0^2$, we must choose n, so that

$$\frac{2ar^n}{(1-r)A_0} < \frac{1}{100}.$$

It is convenient to take for the value I_0 the minimum of the Airy distribution given by the particular phase law $\phi_n = n\psi$. It will be shown that the real phase law on the wedge is not very different from this, and that the intensity maxima obtained under the conditions studied are much greater than this particular value chosen for I_0. Therefore the above condition for n means that intensities in the neighbourhood of the maxima are correct to an accuracy much better than 1%.

To fulfil this condition,

$$r^n = \frac{1}{100}\cdot\frac{1-r}{2(1+r)},$$

we must take n greater than the following values

r	0·93	0·90	0·85
n	115	70	35

Returning to the summation of equation (1), we may take the finite number of terms as above. The same applies to the expansion of (2) in powers of ϵ; in addition we may conveniently neglect terms involving ϵ^4 and higher powers. Let the thickness of the wedge at the position $(0, Y_0)$ be t. Consider first the case of normal incidence ($\theta=0$). Then we have immediately

$$\delta_n = 2nt\left[1-\frac{2n^2+1}{3}\epsilon^2\right] - 2n^2\epsilon^2 X_0. \qquad ......(3)$$

This formula is a good approximation. The absolute error in δ_n because of the neglected powers of ϵ is less than $\frac{2}{3}n^4\epsilon^4X_0$. For the representative case $\epsilon=1$ minute of arc, $X_0=100$ cm., the error amounts to 5000 A. in the hundredth beam.

Equation (2) shows that the intensity distribution on the wedge ($X_0=0$) is not the same as that in Fabry-Perot fringes. For interference of low order (say less than 10) the difference is insignificant. Much of the work with this type of fringe has been carried out with such small gaps.

§ 3. FRINGES NEAR THE WEDGE

A. *Normal incidence*

From (3) we derive the law for the phase of the nth beam, i.e. where $\psi=\frac{4\pi t}{\lambda}$:

$$\phi_n=n\psi-n^2\frac{4\pi\epsilon^2X_0}{\lambda}-n^3\frac{2\psi\epsilon^2}{3}. \qquad \ldots\ldots(4)$$

Near a thickness on the wedge of $N\frac{\lambda}{2}$ we may replace ψ by $(2N\pi+\alpha)$, α being less than 2π, and write (4) as

$$\phi_n=2nN\pi+n\alpha-n^3\frac{2\epsilon^2}{3}(2N\pi+\alpha).$$

If we take two thicknesses differing by $\lambda/2$, the phase of the nth beam relative to the direct is in each case given by this formula. Similar expressions are only obtained if a term $\frac{4}{3}n^3\pi\epsilon^2$ can be neglected in comparison with 2π. Substitution of typical values shows that this is in general the case. The insignificance of this term shows that adjacent fringes always have nearly the same intensity distribution whatever their order.

Further, if the above term is negligible, then $\frac{2}{3}n^3\epsilon^2\alpha$ is also negligible compared with 2π. Hence the phase law becomes very nearly

$$\phi_n=n\alpha-n^3\frac{4\pi N\epsilon^2}{3}.$$

This shows that the intensity distribution changes slowly and periodically, and for the order of interference N it can be calculated as the square of the modulus of the following expression:

$$Ae^{i\Delta}=\sum_n ar^ne^{i\left(n\alpha-n^3\frac{4\pi N\epsilon^2}{8}\right)}=\sum_n ar^ne^{i(n\alpha-\beta\ )}.$$

It is seen from this that the total energy contained in a fringe cycle of 2π is not dependent on the phase relation β_n of the different component vibrations. Hence the main maxima of intensity of a fringe system are all greater than the miminum value of a system obeying the above law with $\beta_n=0$ (i.e. a Fabry-Perot type distribution) as assumed above.

If the phase terms β_n are small, the fringes are asymmetrical. As they become greater, secondary maxima occur in the interference pattern close to the main maxima. In all cases the main period is $\lambda/2$.

B. *Non-normal incidence*

Now consider the case of non-normal incidence. Expression (2) shows that a rotation θ of the axis results in another formula identical with the one for normal incidence. It is therefore obvious that the interference pattern in the plane π_0 for incidence θ is the same as that on the wedge surface for normal incidence. π_0

constitutes the Feussner *surface of localization* (Feussner, 1927). For large values of the thickness, π_0 can be at a great distance from the face of the wedge when the incidence is not critically normal.

It is easily shown experimentally that rotation of the interferometer about the ridge of the wedge, without changing any other part of the system, leaves the appearance of the fringes unaltered. At rotations greater than 20° the fringes split into two polarized components owing to the differential phase change at reflection with angle (Tolansky, 1944).

§4. ASYMMETRICAL BROADENING OF FIZEAU FRINGES

Expanding formula (2) in powers of ϵ for the case $\theta=0$ we find that on the wedge, using similar approximations,

$$\delta_n = 2nt\cos\theta\left(1 - n\epsilon\tan\theta - \frac{2n^2+1}{3}\,\epsilon^2\right). \qquad \ldots\ldots(4)$$

This is the equivalent of the familiar formula quoted for the special case of maxima ($\delta_n = N\lambda$),

$$2t\cos\theta = N\lambda.$$

It is clear from expression (4) that the fringes on the wedge are sharper for normal than for non-normal incidence. The divergence of the phase law from that of Fabry-Perot fringes is, in the former case, $\frac{2}{3}n^2\epsilon^2$, and in the latter it is $n\epsilon\tan\theta$.

Consider the case of normal incidence with a pin-hole source of radius θ. When θ is of the same order of magnitude as ϵ, then any fringe-broadening due to the source extension arises from the $\cos\theta$ term in expression (4) and is significant only for large thicknesses. In effect the broadening is due to the different fringe spacing for each radial zone of the source about the axis. If $\Delta i/i$ is the fractional order broadening at the order of interference N which can be tolerated, the angular diameter of the source must be less than

$$\sqrt{\left(\frac{8}{N}\cdot\frac{\Delta i}{i}\right)}.$$

When the interferometer has a gap of some millimetres, the order of interference is so high that even very small source extensions show broadening, and this broadening appears as a square-topped intensity distribution (figure 3, II).

For gaps less than, say, 100 wave-lengths the same value of $\Delta i/i$ is obtained only for very large extensions (θ) of the source (up to 40°). The broadening is then partly due to the $\cos\theta$ term of equation (4) and partly due to the term $n\epsilon\tan\theta$. The fringes formed by a radial zone θ from the axis are not only displaced towards the side of greater gap but are also broader. Low-order fringes taken with poorly collimated incident light (i.e. non-parallel) exhibit a wing on the side of greater gap (figure 3, III).

This characteristic has found application in the interpretation of surface relief. Figure 4 shows fringes formed in the air film between an optical flat and a sheet of mica, both silvered, and pressed close together. In the first case (figure 4 *a*) strict collimation conditions are obeyed and the fringes have an intensity distribution similar to Fabry-Perot fringes (figure 3, I). When the pin-hole is replaced by an

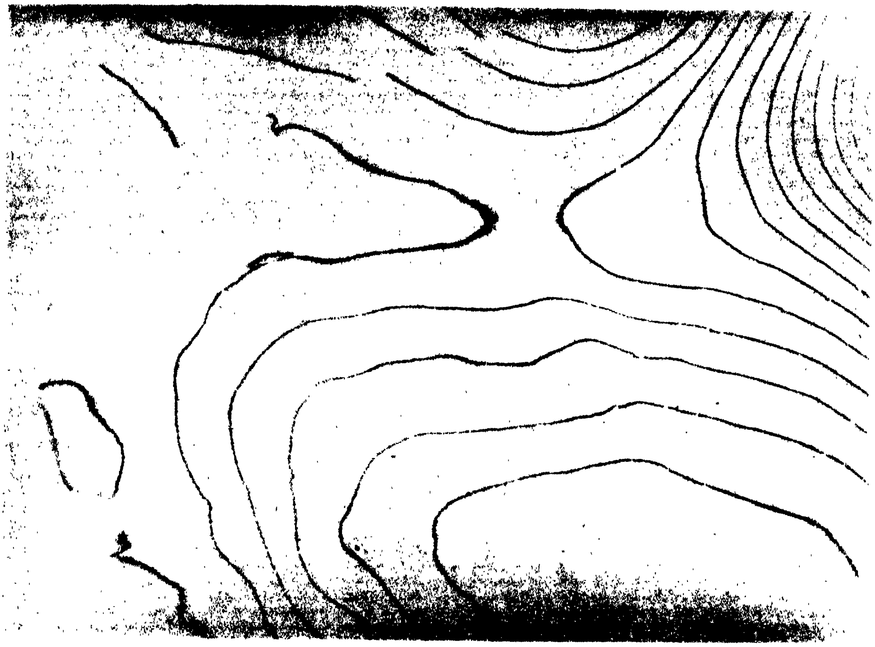


Figure 4 *a*.

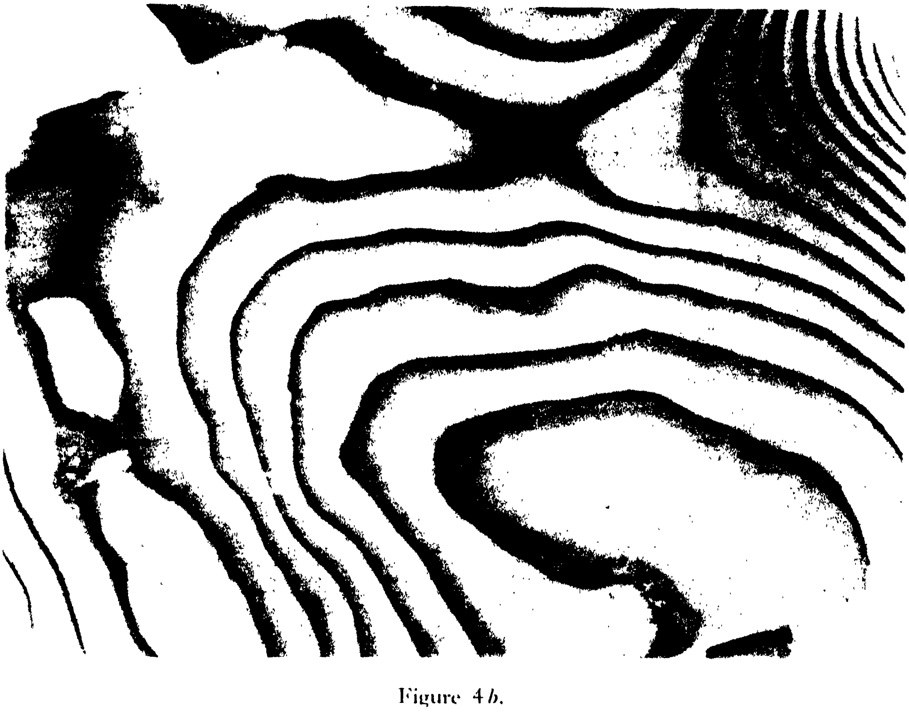


Figure 4 *b*.

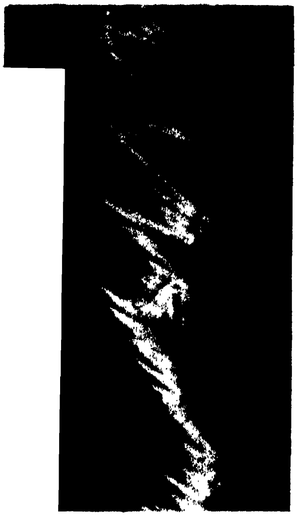
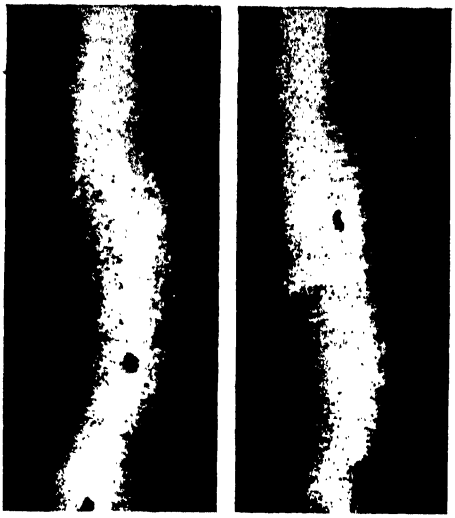


Figure 5.



(a) (b)

Figure 6.

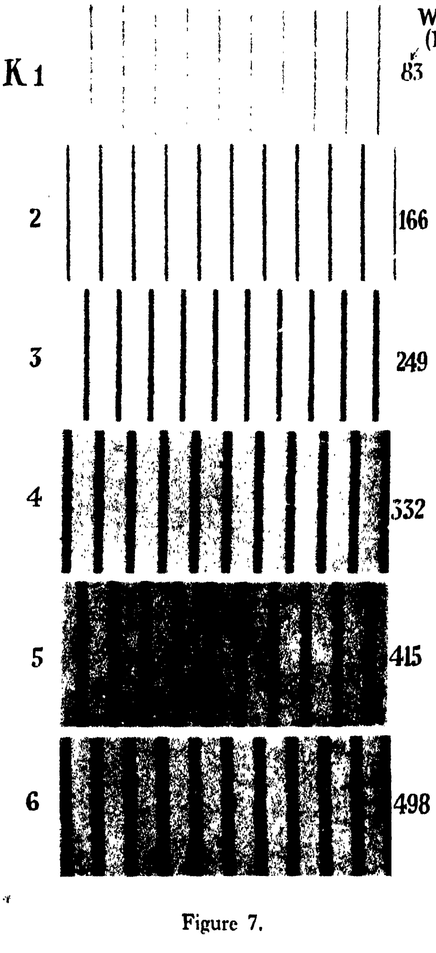


Figure 7.

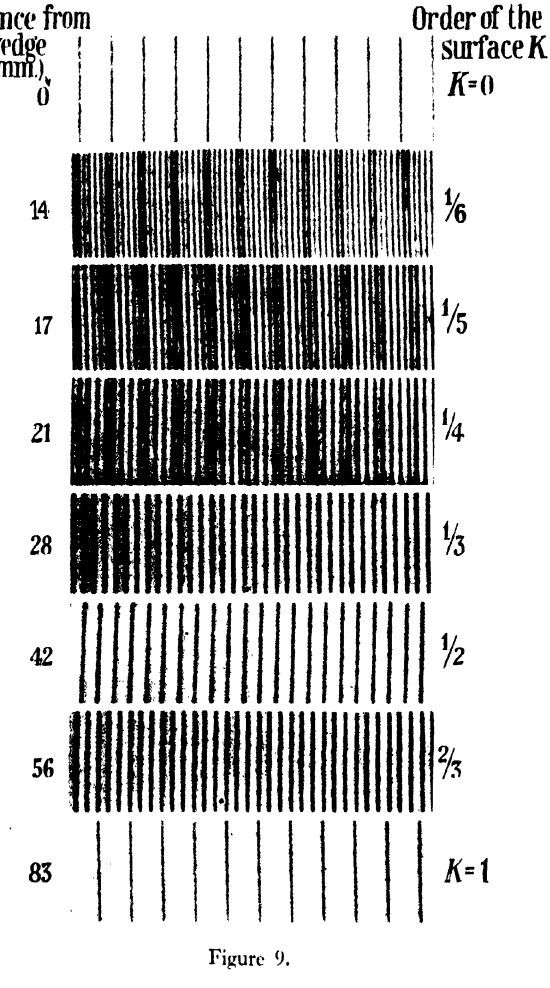


Figure 9.

extended source the asymmetry of figure 3, III appears and the hills and valleys of the surface are clearly differentiated (figure 4*b*).

Any lack of parallelism in the beam results in an asymmetry of this type.

Wings are sometimes observed which seem to be attributable to diffraction and scatter at surface scratches; this is an unavoidable type of collimation error. Good flats usually have a perfect surface polish, and Fizeau fringes obtained between a pair, in a well collimated beam, are perfectly symmetrical and continuous. When one flat is badly scratched the fringes can never be obtained without a certain amount of asymmetry. A good example is shown in figure 5—a photograph kindly supplied by Mr. W. L. Wilcock of these laboratories. The fringes are formed between a polished diamond surface and an optical flat. This photograph was taken under most careful collimation conditions, but scratches on the diamond have caused an asymmetry of the fringes, the side of the fringe towards increasing gap being diffuse. A peculiar phenomenon observed with scratched surfaces is that the fringes with large wedge angles may occur as discontinuous series of dots. The effects are the same with fringes of equal chromatic order, which exhibit the wing on the blue side of the fringe.

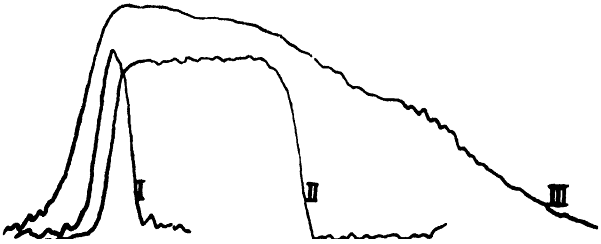


Figure 3.

Figure 6 shows a similar phenomenon with a very small wedge angle. The width of the fringe in the photograph is two per cent of the distance between orders.

In (*a*) the viewing microscope is focused on the plane of the silver. Surface scratches and " pin-holes " in the silver give bright points in the photograph. The same fringe is shown in (*b*) with the microscope at a slightly different focus position. Many of the bright points are now replaced by dark patches surrounded by white diffraction haloes, clearly seen on the side of greater gap (right-hand side). The changes of appearance depend very critically on the position of the microscope, and the size of the details so revealed bears no relation to the real size.

It is of interest to point out that, of the crystal cleavage surfaces so far studied in this laboratory, mica has always given perfectly smooth symmetrical fringes while diamond, calcite, selenite and baryta surfaces have exhibited the characteristics mentioned above. Possibly the effect is due to the existence of an unresolved cleavage structure.

A further cause of fringe asymmetry, arising from light unavoidably off collimation, is the scatter and unwanted reflections from the surfaces of all the optical components. This leads only to very weak wings, of intensity comparable with the normal background between maxima. This has had practical importance only when studying secondary fringe structure.

§ 5. INTENSITY DISTRIBUTION AWAY FROM THE WEDGE

Formula (3) explains an experimentally observed peculiarity of Fizeau fringes. If the collimator has a small pin-hole or a small slit and the interferometer is a simple wedge with its ridge parallel to this slit, then sharp fringes can be observed in many planes other than the Feussner surface.

A. *The interference pattern along the X-axis*

In normal incidence the phase, as shown above, is sufficiently nearly

$$\phi_n = n\psi - n^2 \frac{4\pi\epsilon^2 X_0}{\lambda} - n^3 \frac{2\psi\epsilon^2}{3}. \qquad \ldots\ldots(4)$$

This formula also applies to non-normal incidence if X_0 is replaced by x_0, which is the rotated axis perpendicular to the incident wave front.

Taking two planes parallel to π_0 and a distance apart of

$$x = \mu \frac{\lambda}{2\epsilon^2} \quad (\mu = 0, 1, 2, \ldots) \qquad \ldots\ldots(5)$$

we see that the families of component vibrations arriving at these planes differ in phase, member by member, by $n^2\mu \,.\, 2\pi$. It follows that in any two planes parallel to the incident wave-front and a distance apart defined by (5), the interference patterns are exactly the same.

B. *Recurrence of fringes similar to those on the Feussner surface*

In particular, if the distances are taken from the Feussner surface then the observed pattern in each plane is identical with that on π_0 itself.

In a series of experiments designed to observe as many as possible of the orders (μ) of the surfaces of localization, two silvered flats were used. A large value of ϵ was adopted, there being 205 fringes in a field of diameter 43·2 millimetres with the mercury line 5461 A.

Formula (5) gives, from these figures, the value $x = 16{\cdot}5$ cm.

When a microscope was focused first on the fringes localized on the flats themselves and then successively on the planes x, $2x$, $3x$, ... $10x$, closely similar fringe systems were observed, the appearance of the field being quite different away from these positions. The only exception was that exactly in the middle of the above locations a similar pattern again occurred (i.e. at 8·2 cm., 16·5 cm., etc.). This property is general because on planes distant $\frac{x}{2} = \frac{\lambda}{4\epsilon^2}$ apart the interference patterns are the same, apart from a half-order displacement.

The proof is as follows:

Let ϕ_{1n} be the phase relation on the first plane and ϕ_{2n} be that on the second. We have

$$\phi_{2n} = \phi_{1n} - n^2\pi.$$

Now change the origin of ψ in equation (4) by π for the first surface only. With this new origin we find the phase relation on this surface becomes $\phi'_{1n} = \phi_{1n} - n\pi$ if a term $2\pi n^3\epsilon^2/3$ is neglected (as was done before in showing that adjacent fringes in a pattern have practically the same intensity distribution). Obviously ϕ_{2n} and ϕ' are identical because n and n^2 are even or odd together.

This half-order fringe shift is difficult to demonstrate experimentally. In the present case it was 0·1 mm. for positions of the microscope 10 cm. apart. In practice, therefore, it is convenient to consider that the interference repeats at intervals of $K\frac{\lambda}{4\epsilon^2}$, and these planes of repetition may conveniently be designated as the surface of localization of order K, say F_K. Figure 7 shows photographs of the patterns observed in the planes where K is integral, the exact values being indicated in the figure. Displacements of half an order have been given to the even-number patterns to accord with the above theory.

Patterns with spacing different from this were observed between these positions and will be described below. The effect of the extension of the source will be considered first.

If a perfect point source in the focal plane of the collimator is moved slightly off the axis in a direction parallel to the ridge of the wedge, the fringes in F_K are merely displaced parallel to their length and the appearance remains unaltered.

If the point source moves so that the incident plane wave front π_0 rotates through $d\theta$, remaining parallel to the ridge of the wedge, the fringes in F_K rotate about the apex of the wedge through the angle $d\theta$, their linear displacement then being $x_K \, . \, d\theta$.

The photographs in figure 7 were taken with a small slit source in the collimator arranged with its axis parallel to the ridge of the wedge. The square-topped intensity distribution is easily explained on the grounds of the above considerations. On a given surface F_K the fringe width is proportional to the source width $d\theta$, and for a given source it is proportional to the order K of the surface. The latter effect is apparent in figure 7.

C. *Differently spaced fringes in other planes of localization*

Between the surfaces with integral values of K, other positions are found where well-defined fringe systems occur. For example, we may take the half-way positions. The intensity is given by the square of the modulus of

$$A_{\frac{1}{2}} e^{i\Delta_{\frac{1}{2}}} = \sum_n a r^n e^{i\left(n\psi - n^2\frac{\pi}{2}\right)}.$$

Taking the real part of this expression, the summation becomes

$$A_{\frac{1}{2}} \cos \Delta_{\frac{1}{2}} = \sum_p a r^{2p} \cos p(2\psi) + \sum_p a r^{2p+1} \sin (2p+1)\psi.$$

Each of these series can be calculated from the trigonometrical formula

$$\cos\eta + \rho\cos(\eta+\zeta) + \ldots + \rho^n \cos(\eta + n\zeta) + \ldots = \frac{\cos\eta - \rho\cos(\eta-\zeta)}{1 - 2\rho\cos\zeta + \rho^2}.$$

Finally, for the real part,

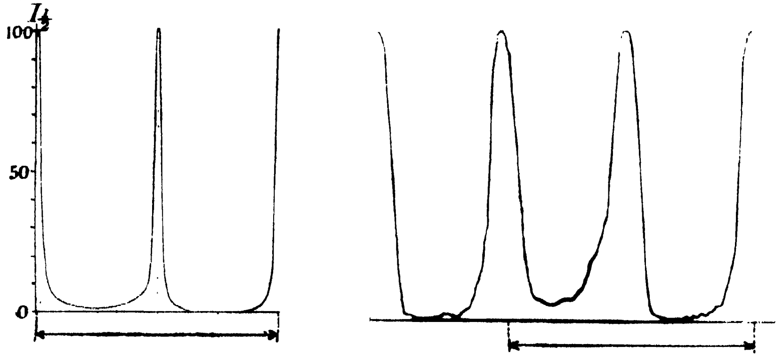
$$A_{\frac{1}{2}} \cos \Delta_2 = a\,\frac{1 - r^2\cos 2\psi + r(1+r^2)\sin\psi}{1 - 2r^2\cos 2\psi + r^4}.$$

A similar calculation applies to the imaginary part, and the following table gives the resulting numerical intensity values for $r = 0{\cdot}92$. Figure 8 (*a*) gives the curve, and figure 8 (*b*) is a direct photometer record taken across a photograph of the fringes.

ψ	0°	5°	10°	20°	90°	180° $-\xi$	180°	180° $+\xi$	190°	200°	210°		270°		360° $-\xi$	360°
$I_{\frac{1}{2}}$	100	53	21	7·2	0·32	53	100	46·5	14	3·6	1·3	0	$2\cdot10^{-3}$	0	46·5	100

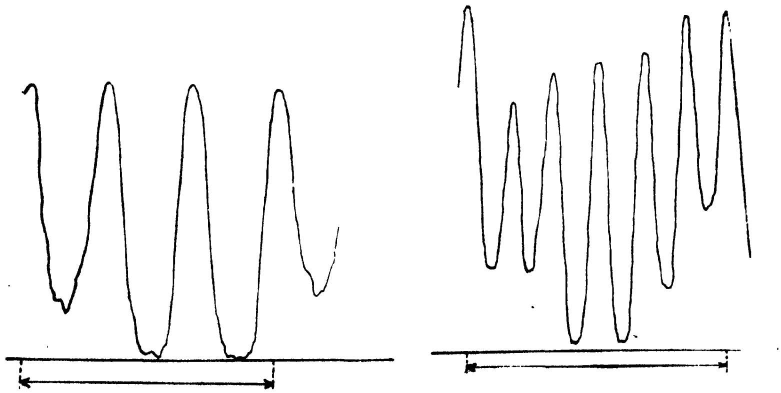
$$\xi=\frac{1-r^2}{2r}\quad\text{and for } r=0{\cdot}92,\ \xi\approx5^\circ.$$

Sharp maxima occur with a spacing half that of the fringes on the Feussner surface. The minima, however, alternate in intensity so that the true period of these fringes is the same as the period of those on the wedge. Attention is drawn to the very low intensity of one background compared with the Airy distribution.



(*a*) $K=\frac{1}{2}$ calculated.

(*b*) $K=\frac{1}{2}$ record.



(*c*) $K=\frac{1}{3}$ or $\frac{2}{3}$.

(*d*) $K=\frac{1}{6}$.

Figure 8.

A similar calculation gives the intensity distribution at intervals $\frac{1}{3}$ and $\frac{2}{3}$ of the interval between the surfaces $\mathrm{F_K}$ (K integral). The fringes obtained have one-third of the spacing between maxima, and every third background is relatively intense (figure 8(*c*)). Figure 9 shows other intermediate cases at K values of $\frac{1}{4}$, $\frac{1}{5}$, $\frac{1}{6}$. All these photographs were over-exposed to record the background intensities. The microphotometer record for $K=\frac{1}{8}$ (figure 8(*d*)) shows that the maxima do not have equal intensity, two adjacent maxima of each cycle being quite intense, and exhibiting a high intermediate background. As the wedge is approached from this position (i.e. $K\rightarrow 0$) this characteristic becomes more dominant, until the Airy distribution is reached.

The intensity distribution can be calculated in the more general case of $K=1/m$ (m being any integer).

Assuming the phase law $\phi_n = n\psi - n^2\frac{\pi}{m}$, the intensity is given by the square of the modulus of

$$A_{1/m}e^{i\Delta_m} = 1 + \sum_{q=1}^{q=m}\left\{\left(re^{i\Psi}e^{-iq\frac{\pi}{m}}\right)^q \frac{1}{1-(-1)^m r^m e^{im\Psi}}\right..$$

This phenomenon of multiple localization occurs with many types of multiple-beam interferometer, and recalls the properties of the zone plate. In fact the experimental arrangement of the wedge described above is similar to Fresnel's mirror or the Billet split lens, being merely the multiple-beam version of it. With the two beam interferometers. the fringes always have a sin² intensity distribution and the order spacing is always given by the simple formula $\delta = N\lambda$. For the case of multiple beams, the fringes appear as very sharp lines in most cases, but the distance between the lines is given by the above formula in only a few instances. If, however, it is borne in mind that the true fringe cycle is determined by secondary phenomena (fringe background and inequality of maxima) it is easily seen that the true period of the fringes always obeys the relation $\delta = N\lambda$.

D. *Recurrence of the interference pattern along the wedge*

We have just examined some of the consequences of the phase law

$$\phi_n = n\psi - n^2\frac{4\pi\epsilon^2X_0}{\lambda},$$

in which

$$\frac{4\pi\epsilon^2X_0}{\lambda} = K\,.\,2\pi;\ \pi;\ \frac{\pi}{2}.$$

On the wedge itself the phase law is more exactly given by

$$\phi_n = n\psi - n^3\frac{8\pi t\epsilon^2}{3\lambda},$$

as seen from equation (4). This term in n^3 gives the same effects as noted above for the n^2 term when it is equal to $2N\pi$, π, $\pi/2$, because n^2 and n^3 are even and odd together and the amplitude series are identical for the two cases.

Hence at the thicknesses of the wedge given by

$$t = N\,.\,\frac{3\lambda}{8\epsilon^2}$$

a Fabry-Perot type intensity distribution should exist.

An interesting deduction is that between two thicknesses differing by $3\lambda/4\epsilon^2$ the number of fringes observed should be one less than that predicted by the simple theory.

Tolansky (1943) has noted a similar deviation of fringe spacing from the predictions of simple theory in the case of non-localized Fabry-Perot rings (Tolansky, 1943 b).

Another important point is that at thicknesses

$$t=\frac{3\lambda}{16\epsilon^2}+N\frac{3\lambda}{8\epsilon^2}$$

fringes should be observed which are bright lines on a dark background and have a distance between them only half that given by the simple theory.

It is proposed to investigate this prediction experimentally in this laboratory in the near future.

ACKNOWLEDGMENTS

I wish to express my thanks to Dr. S. Tolansky for his interest and encouragement. Among my co-workers, I am particularly indebted to W. K. Donaldson, who first constructed much of the apparatus necessary for this research.

REFERENCES

Feussner, 1927. Gehrcke's *Handbook der Physik—Optik*, vol. i.

Tolansky, 1943 a. *Nature, Lond.*, **152**, 722; 1943 b. *Phil. Mag.*, **34**, 55; 1944. *Ibid.*, **35**, 120; 1945. *Ibid.*, **36**, 225.

MULTIPLE-BEAM LOCALIZED FRINGES: PART II.—CONDITIONS OF OBSERVATION AND FORMATION OF GHOSTS

By J. BROSSEL,
Paris
(Now at Manchester University)

Communicated by S. Tolansky; MS. received 4 *September* 1946

ABSTRACT. Factors which tend to reduce the perfection of the fringes exist in the optical systems associated with a Fizeau multiple-beam interferometer and in its complexity of surface detail. A parallel is drawn between the mechanism of image formation of these fringes and the theory of microscope image-formation due to Abbe. Experiments are described showing the effect of the presence of diaphragms in the second focal plane of the objective and also in contact with the objective. The conclusion reached experimentally is that if the objective has insufficient aperture the fringes broaden and secondary maxima appear.

At the same time, the resultant of a non-infinite series of geometrically decreasing waves is studied. The results obtained are also valid for the Lummer-plate interferometer.

Detailed attention is given to the parasitic reflected light from the many glass-air surfaces of the optical train used, and the resulting ghosts are described.

It is shown that the interferometer gap must be of only a few wave-lengths if the localized fringes are to follow the contour lines correctly over the smallest surface details.

§ 1. INTRODUCTION

In this paper an examintion is made of the necessary conditions which must be fulfilled in the observation of multiple-beam Fizeau fringes in order to obtain a final image identical with the primary interference pattern. The investigation was undertaken when the question arose as to the limit of surface detail which can be resolved by the technique, and the accuracy with which the fringes follow surface contours.

It is assumed throughout that the order of interference is very small, that the idealized case of a simple wedge is under consideration, and that the fringes are those localized in the Feussner surface of zero order.

§ 2. EFFECTS OF THE PRESENCE OF DIAPHRAGMS ON THE STRUCTURE OF THE IMAGE

(a) *Abbe type experiment*

The present observations resemble the phenomena studied by Abbe in developing his theory of the resolving power of the microscope and, in particular, his case of a grating (R) in a parallel beam (figure 1 (*b*)) and the conditions for similarity of the image to the object. Comparing the optical arrangement used here (figure 1 (*a*)) with that of Abbe, it is apparent that the only difference lies in the method of division of the wave. Abbe has clearly shown that the image E of the grating R is an interference pattern arising from the reciprocal grating R_1 in the focal plane of the objective. Artificial modifications of R_1 produce false detail in the image.

In the interferometer case there is division of amplitude of the incident beam, and the emergent family of multiple reflected wave fronts $\pi_0, \pi_1, \ldots, \pi_n$ (see Part I; Brossel, 1947) give a set of equally spaced images of the original point source in the focal plane of the microscope objective M. The intensities of these images decrease geometrically, and there is a definite phase relation between them. For convenience these will be referred to as the grating I_1, reciprocal to the interference phenomenon I. The fringes formed on E can be calculated from this grating I_1: any defect of phase or amplitude imposed on the grating I_1 will cause fundamental modification of the final image of the fringes.

The functioning of many multiple-beam interferometers can be explained in terms of this grating—in particular the interferometers of Fabry-Perot and Lummer-Gehrcke. In Part I it was indicated that Fizeau fringes under discussion are the multiple-beam version of the two-beam fringes of Fresnel's mirrors or Billet's split lens. The relationship is clearly in evidence from figure 1. These fringes, E, could be considered, too, as the different orders of spectra coming from an echelon. The plate thickness would be equal to the thickness of the interferometer on the optical axis, but the angular dispersion would not be due to diffraction (there is no amplitude factor dependent on direction).

These multiply-reflected images of a point source were used by Fabry (1927) in the determination of reflection coefficients. They served the same purpose in this investigation, but a photographic method was adopted in place of his polarization technique. Their use more generally in photographic photometry has been overlooked in the past, but they form the only true logarithmic intensity scale. The source does not require stabilizing, but the reflectivity usually has to be evaluated by some auxiliary method.

The phase relationship between the elements of the grating I_1 can be calculated by a simple geometrical method. Translation of the interferometer leaves the positions of the images I_1 unchanged, but modifies their phase relationship. When the wedge is in the focal plane of the objective, with its apex on the axis, all the sources I_1 are in phase and the resultant final fringes at infinity are the maxima of a grating. A translation in the plane of the figure perpendicular to the optical axis introduces a constant phase difference between adjacent sources which ensures the corresponding translation of the final image pattern.

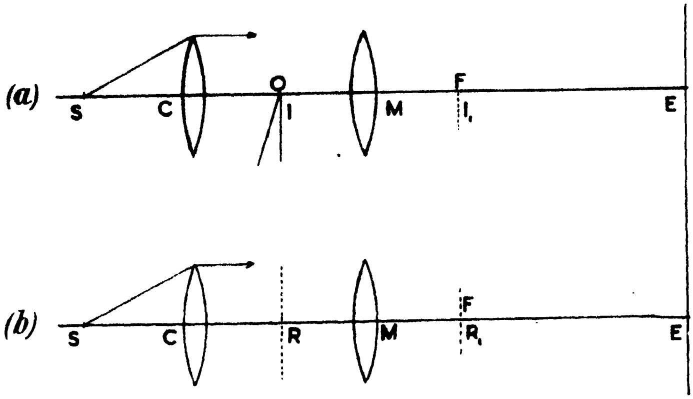
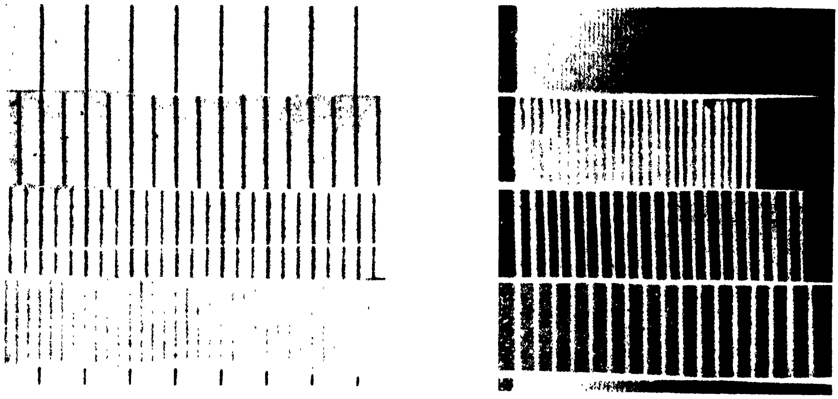


Figure 1.

When the interferometer I is at a distance $\overline{MO} = u$ from the objective, the path difference between the first and the nth source is $2n^2\epsilon^2(f-u)$. (The notation of Part I is used throughout.) It is possible to choose the experimental conditions so that the supplementary paths involved exactly compensate this value. It is sufficient for this purpose to place the screen E so that $\overline{ME} = v = \frac{uf}{u-f}$ i.e. in the plane conjugate to I. The image of the fringes can be very close to the grating I_1, from which they arise, and yet remain perfect.

In planes other than this conjugate, the multiple localization phenomena of Part I are observed, but it is apparent that they may also be considered as arising from the grating I_1.

Consider now the effect on the fringes at E, arising when various obscuring diaphragms are placed in the plane I_1. For instance, alternate images can be masked and the resulting grating is then equivalent to that given by a wedge of twice the angle, and having a reflectivity r^2 instead of r. The result is that the fringes observed with such a mask will have a spacing only half that of the

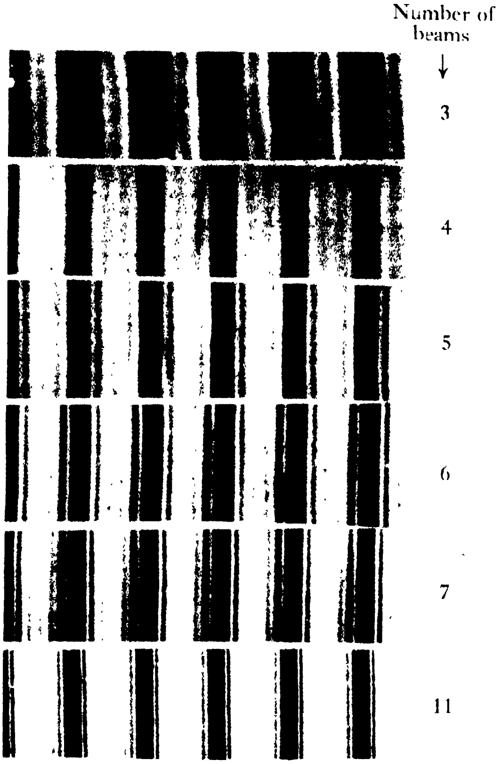


(*a*) (*b*)

Figure 2.

(*a*) Number of beams 30.

(*b*) Number of beams ∞.



(*c*)

Figure 3.

Figure 5.

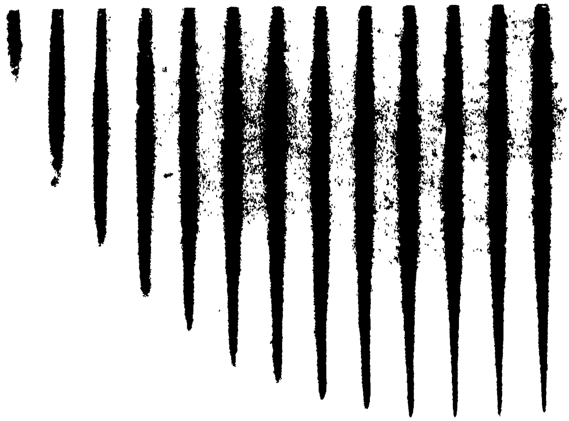


Figure 6.

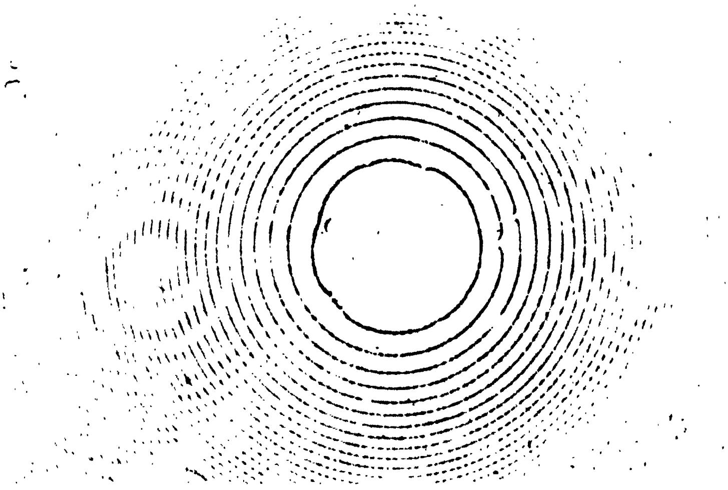


Figure 8.

original fringes, and must be broader. This corresponds exactly with observation, and figure 2 shows the result of a series of experiments of this type.

The masks, shown in figure 2(*b*) (excluding top and bottom pictures), were suitably adjusted in the focal plane I_1 of the objective, and the corresponding fringes observed are shown opposite each in figure 2(*a*). To complete the picture, the top photograph on the left shows the unmodified fringes, and at the same level at (*b*) are shown the unmodified multiple images in plane I_1. (The source was not a pin-hole but a small slit.) When two out of three, three out of four, etc., images are masked in I_1 the observed fringes are three times, four times, etc., as close together.

The multiple images were first recorded on photographic paper which was then suitably cut with a razor blade to make the mask. This was then adjusted in the focal plane I_1 with the aid of a micrometer screw and a high magnification simple eyepiece. The projection objective used to form the fringe image was a Zeiss anastigmat, f/4·5 and 50 cm. focal length.

(b) *Secondary maxima*

A second set of experiments was performed in which the series of images was restricted to a given number, the less intense tail of the diminishing series being masked. This is a convenient method of studying the combination of a non-infinite series of waves of diminishing amplitude. This case is met with in practice when the objective has an insufficient aperture, as detailed below. The results also apply to the Lummer plate, which seldom uses more than 30 beams. If the grating I_1 is restricted so that only the first two sources are uncovered by the mask over the series tail, the fringe images have a $\sin^2$ intensity distribution. For three sources (the three-line grating) the principal maxima are sharper than in the previous case and there is one secondary maximum between successive orders. With four sources there are two secondary maxima, and so on. The development of the phenomenon was followed in detail with an increasing number of sources. The number of secondary maxima is identical with that in the classical grating case of Abbe, but the intensities are influenced by the fact that in the present case a logarithmic decrease of intensity obtains. With a reflection coefficient of 0·92 and 20 sources the 18 secondary maxima were easily photographed. Beyond this the secondary maxima half-way between principal maxima were lost in the background, but the secondaries close to the principal maxima were resolved when as many as 40 sources were used. Figure 3(*a*) shows a photograph taken with 30 sources, and four or five secondaries are apparent. Figure 3(*b*) shows the same fringes with an "infinite" series of sources (Fabry-Perot distribution). A number of intermediate cases are given in figure 3(*c*), the number of exposed sources being printed to the right of each photograph. (The fringes are black.)

The intensity distribution in the fringes given by 30 sources has been studied by photographic photometry and the curve obtained is given in figure 4. These measurements apply to an order of interference of 25. The term $\frac{4}{3}n^3\epsilon^2 t$ is important enough to cause a distinct asymmetry (see Part I). On the scale given, the peak of the maximum (omitted) has intensity 100.

As a result of this it appears that a multiple-beam interferometer should utilize at least 40 beams. For instance, the length of a Lummer plate should be at least 80 times its thickness. If this condition is not fulfilled, great care should be taken when the intensity and position of a hyperfine structure component close to a strong maximum and less than 10 % of its intensity are being measured. These data apply to a perfect instrument used with very sharp lines, and the tolerances are often greater in practical cases.

(c) *Effect of the aperture of the objective.*

Returning to the conditions desirable for observing Fizeau fringes, it is immediately apparent from the above that the aperture of the projecting lens which collects the beams is a fundamental consideration. When this is insufficient it acts as a mask and the fringe image is broadened, whilst secondary maxima appear. The beams leaving the interferometer all fall on the same side of the optical axis. If the lens diameter is less than the illuminated field of the interferometer, all the beams (or, say, 100) on one edge of the interferometer can pass through the objective: the corresponding fringe image will have the Fabry-Perot distribution. At the opposite edge of the interferometer field only one or two of the beams leaving the interferometer pass through the objective: the corresponding fringe image will then have a sine-squared distribution. Between these two extremes all the intensity distributions described above are met with; in traversing the field from the broad fringes, the principal maxima become sharper and sharper, the secondary maxima increasing in number. Finally, these secondaries are lost in the background and the region of sharp fringes is reached.

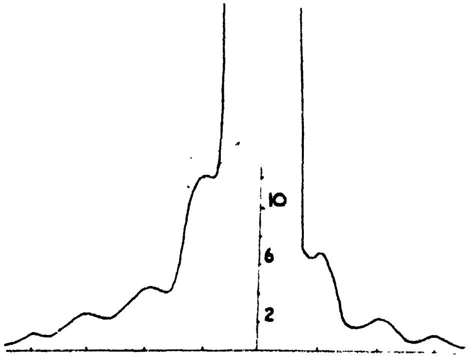


Figure 4.

A photograph of this effect is given in figure 5. The diaphragm across the objective was a slit 4 cm. wide, with its edges parallel to the fringes. The field of the interferometer was 6 cm. in diameter.

Similar effects are also observed with a circular diaphragm, but they differ in that the intensity distribution may vary along the length of a fringe. The photograph in figure 6 was taken under similar conditions to that of figure 5, the rectangular objective stop being replaced by a circular hole of 4 cm. diameter.

In the microscope normally used to observe Fizeau fringes the field diaphragm is small enough to prevent such an effect. Occasionally the fringe definition at one edge of the field is observed to be imperfect.

An alternative to modifying the amplitudes of the sources I_1 is the modification of their relative phases.

A random variation in phase is readily imposed by introducing a plate of glass in the plane I_1. The definition of the observed fringes is then materially affected.

If a vessel constructed like a hollow echelon were placed correctly in the plane I_1, the instrument would behave as a refractometer—being in fact the multiple-beam version of the Rayleigh instrument. Of course it is much simpler to introduce the dispersive medium between the plates of the interferometer itself and use gaps of the order, for instance, of centimetres.

§ 3. GHOST IMAGES IN FIZEAU FRINGES

(a) *Glass surfaces before the interferometer*

In studying the secondary maxima it was found that weak systems other than those mentioned above are present. These "ghosts" have a comparable intensity to the secondary maxima of interference—in the region of 4% of the main maxima. They are due to parasitic reflections from the unsilvered glass surfaces. 90% of the light incident on the interferometer is reflected by the first silvered surface and 4% of this is returned towards the observer from each glass surface encountered. Each of these surfaces therefore produces a ghost point-source which has its own set of fringes.

The optical arrangement used is shown in figure 7; each surface is numbered, and a monochromatic filter is usually included (3, 4). The plane surfaces (3 and 4) of the filter and the unsilvered surface (5) of the first optical flat give ghost point-sources s_3, s_4, s_5 at infinity and in general these do not lie on the axis. As a result (because of the $\cos\theta$ law) the corresponding ghost fringes have each a distance between orders slightly greater than the main system, and appear as a fringe structure on the side of greater gap. The resolution of this structure improves with increasing order. These ghosts can be eliminated by increasing the tilt of the offending surfaces, or by fitting a screen in the second focal plane of the objective so that only the main set of multiple images passes.

Surfaces 1 and 2 of the collimating lens give two ghost point-sources s_1, s_2 at finite distances. The ghost fringes are those formed in a wedge with a point-source at a finite distance—a much more complicated situation than the parallel beam analysed above. It can be shown that locally the fringes formed are the same as if the source were moved to infinity along the actual direction of incidence. At the foot of the perpendicular to the wedge dropped from s_1 (or s_2) the ghost and the main fringe coincide, and away from this region in any direction (along the fringe too). the ghost moves to the side of greater gap of the main fringe. This has been observed experimentally.

It is difficult to eliminate the ghosts 1 and 2. The collimating lens may be tilted a little and the interferometer placed at a distance such that the ghost beams are not collected. It seems that the best solution is to use a plano-convex lens with its plane side towards the interferometer. Ghost 2 is then in the same class as 3, 4 and 5 above and the real point-source s_1, if troublesome, can be brought to a focus on the interferometer in an unimportant area.

(b) *Surfaces following the interferometer*

The ghosts arising from surfaces 6, 7 and 8 on the exit side of the interferometer are due to the reflection in each case of 4% of the incident energy, and of this 90% is reflected again by the second silvered surface towards the observer. These ghosts are images of the fringes given by the mirrors 6, 7 and 8 in combination with the plane-silvered mirror and the objective. They are not localized on the interferometer and can therefore exhibit great complexity since the plane of observation is not the true plane of localization. In these experiments the ghost 6 is localized at about 3 cm. behind the main interference pattern. The surface of the microscope objective (7) is plane, and with the usual focal lengths its ghosts fringes are localized at 7 to 10 cm. behind the interferometer. With the small dispersion used (angle of the wedge a few minutes of arc) the pattern away from the surface of the wedge does not change rapidly with X (Part I—Intensity distribution away from the wedge) and ghost 6 always appears simply as an out-of-focus fringe superposed on the main system. Ghost 7 is seldom observed. Surface 8 produces an inverted image of the fringes, which always seem to be localized close to the interferometer. A bloomed objective appears to offer the only improvement. This ghost is easily identified when using Fizeau fringes; as the field of the interferometer is explored under the microscope, the ghost and the main fringe system move in opposite directions.

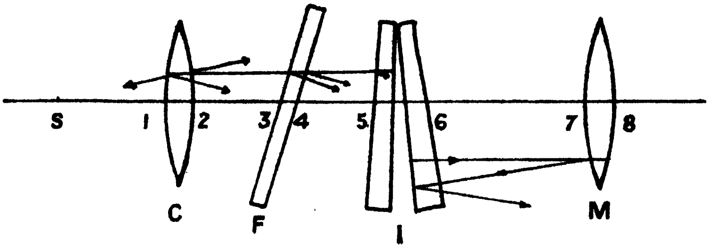


Figure 7.

(c) *False details arising from the ghosts*

These ghosts from the surfaces of the optical components frequently cause false detail to appear in the fringes. It must be remembered that when studying crystal topography with a microscope there can be more unsilvered glass surfaces than those described above, and each gives some form of ghost. The most troublesome type of ghost arises when two surfaces close together are nearly parallel. The result is that the intensity of the main fringes is modulated by the ghost pattern. This effect in multiple-beam Fizeau fringes is well known, having been formerly described, and is mentioned here for the sake of completeness. A photograph clearly demonstrating the effect is given in figure 8, and shows multiple-beam Newton's rings with two $\sin^2$ ghosts modulating the main fringe system. It is rarely that the bright parts of the main fringes can be joined up to give a simple secondary system as here.

§4. EFFECTS OF SURFACE STRUCTURE

In conclusion it is apparent that large-aperture objectives must be used to project Fizeau fringes if a sufficient number of the interfering beams are to be collected. Further, since the objective is collecting only plane or nearly plane waves, it is working under the worst conditions for the resolution of the true details of the surface, as shown by Abbe. This applies to the resolution in directions perpendicular to the axis. The Fizeau method, however, is used only to reveal details in the dimension parallel to the axis. In this dimension the improvement due to the use of multiple beams over other methods is enormous, being comparable with the power of the electron microscope in the other two dimensions.

The details which are revealed are those which introduce a phase difference. At intervals of $\lambda/2$ of optical distance from the surface of the reference flat, plane parallel zones in space may be imagined, of thickness only a few per cent of the distance between zones. If a detail on the surface under investigation is to appear at all in the pattern observed, it must fall within these zones. For the greater part of each zone there is a visible change of intensity between parallel planes distance apart only a few ångströms. Between the narrow zones of high intensity gradient there is practically no light and no intensity gradient, and a step of 2000 A. can escape notice.

When studying a surface, therefore, it is essential that every point of the surface must at some time fall within a zone of maximum light.

Up to this point the conditions to ensure that the final image is similar to the fringe pattern localized in the wedge have been treated. Whether these fringes follow the true contours will now be considered.

In the general case of a surface against an optical flat, a detail is observed either as an area of different tint or as a short displaced part of a fringe. The only occasion when the interpretation of size or contour may be erroneous is when the region is extremely small. The fringe delineates the true contour if the interference conditions are fulfilled over this small region.

For example, the detail observed may be a flat-bottomed channel perpendicular to the fringes and 0·01 mm. wide. The incident plane wave will be diffracted through large angles and so will each of the multiply-reflected waves. If the optical flat is very close to the surface, say five waves (a distance only a quarter of the width of the channel), the diffraction is restricted to the edges, and inside the channel the waves are effectively plane and the interference conditions are fulfilled over most of the width.

An exact calculation of the effect does not appear to be fruitful. Experimentally, many observers in this laboratory agree that the definition and perfection of details revealed is much greater the lower the order of interference. This also depends on the fact that the lateral displacement of the beams down the wedge is more easily kept within the boundaries of the details when very thin gaps are adopted.

One effect of this diffraction at the surface steps is that a local wing is produced at the fringe break on the side of greater gap. It seems likely that the sharp edge of the fringe represents the true contour of the surface for details down to sizes of 0·01mm.; crystalline cleavage zones of this size can be perfectly resolved. Deductions about minute detail are not too reliable, especially when one considers the critical dependence on focus of fringe detail. The fringe reproduced in figure 6 of Part I of this paper is a typical example of this latter effect, the granular structure of the fringe having an appearance dependent entirely on focus.

Many existing conditions for the effective formation and use of multiple-beam Fizeau fringes are favourable: perfection of the surface is desirable over only highly localized regions; many sources of trouble are readily eliminated by the use of a low interference order, the fringes are easily obtained bright and very sharp. The technique is thus a considerable advance on any previous application of the fringes of equal thickness. The measurement of cleavage steps, surface curvature, etc., can be still further improved by adopting fringes of equal chromatic order (Tolansky, 1945) in place of these fringes. Both types of fringe obey the same fundamental formulae, and all the results described above for Fizeau fringes can be adapted to the description of fringes of equal chromatic order.

REFERENCES

Brossel, J., 1947. " Multiple Beam Fizeau Fringes : Part I." *Proc. Phys. Soc.*, **59**, 224.
Fabry, C., 1927. *Les Applications des Interferences Lumineuses* (Paris : Ed. Rev. Opt. théor. instrum.).
Tolansky, S., 1945. *Phil. Mag.*, **36**, 225.

THE DETERMINATION OF THERMAL LAGGING TIMES

By J. C. EVANS,
The National Physical Laboratory

MS. received 15 *September* 1946

ABSTRACT. The problem discussed in this paper arose out of the need for accurately determining the temperature of a barometer under conditions of varying air temperature. A method for computing the thermal lag of simple and complex bodies of cylindrical form, when the ambient temperature is changing steadily, is given and formulae derived for determining the lagging times of such bodies in liquid and in gaseous media. Values calculated from the formulae are compared with results obtained experimentally.

§ 1. INTRODUCTION

The author was led to consider the question discussed in this paper when engaged, some years ago, in investigating the degree of precision attainable in the measurement of pressure by means of the mercury barometer of normal design. The theory which was developed has ever since served as the basis of temperature measurement in barometric work at the National Physical Laboratory and has proved fully satisfactory. For this reason, and because it may be found of use in other applications, it seems desirable to place the theory on record.

The temperature coefficient of a Fortin barometer of normal design is large, a change of temperature of 1° c. corresponding to a change of 0·12 mm. in the reading when the pressure is constant and of the order of 760 mm. of mercury. In precision measurements, pressures are required to 0·01 mm. or better, and a determination of the true temperature of the barometric column is therefore essential. As ordinarily supplied, the barometer is fitted with a thermometer mounted on the brass sheath, but with its bulb between the sheath and the mercury tube. Under conditions of varying air temperature, this arrangement clearly gives only the approximate temperature of the barometric column, and for more specialized applications other methods have been introduced. For example, the thermometer is sometimes mounted with its bulb immersed in the mercury in the cistern of the barometer; alternatively, the thermometer is mounted alongside the barometer, its bulb immersed in mercury held in a glass tube of the same bore as the barometer tube. Experience has shown that in the former case the temperature of the thermometer lags behind that of the barometric column, whilst in the latter the temperature of the column lags behind the thermometer.

§ 2. THEORETICAL BASIS OF CORRECTION

An allowance for the error which results from the use of an incorrectly matched thermometer may be made if the "lagging times" of the column and of the thermometer can be determined; alternatively, the lagging times can be made the same by providing a suitably shielded thermometer. The definition of lagging time emerges from the following considerations.:—

(1) Assume a body at temperature θ_0 immersed at time $t=0$ in a medium at the higher temperature ϕ. The initial rate of change of temperature of the body, assuming Newton's Law, is proportional to the temperature difference $\phi-\theta_0$.

If θ is the temperature at time t,

$$\frac{d\theta}{dt}=\frac{1}{\lambda}(\phi-\theta),$$

in which λ is a constant, and therefore

$$\phi-\theta=(\phi-\theta_0)e^{-\frac{t}{\lambda}}. \qquad \ldots\ldots(1)$$

The constant λ is known as the lagging time of the body in the medium postulated. After the lapse of an interval λ the difference between the temperature of the body and that of the medium will have been reduced to $1/e$ of its initial value.

(2) Assume that a body at temperature θ_0 is immersed in a medium at the same temperature and that at time $t=0$ the temperature of the medium commences to rise uniformly at the rate K per unit time.

Then, at time t,

$$\frac{d\theta}{dt}=\frac{1}{\lambda}(\theta_0+Kt-\theta),$$

which gives

$$\theta_0+Kt-K\lambda-\theta=Ae^{-\frac{t}{\lambda}}$$

or, since

$$\theta=\theta_0 \quad \text{when} \quad t=0,$$

$$\theta_0+Kt-K\lambda-\theta=-K\lambda e^{-\frac{t}{\lambda}}.$$

Hence and

$$\left.\begin{aligned}\theta&=\theta_0+Kt-K\lambda(1-e^{-\frac{t}{\lambda}})\\ \frac{d\theta}{dt}&=K(1-e^{-\frac{t}{\lambda}}).\end{aligned}\right\} \qquad \ldots\ldots(2)$$

After the lapse of a period of time which is large compared with λ, the rate of change of temperature of the body is identical with that of the medium, but the temperature of the body lags behind that of the medium by the amount $K\lambda$, i.e. by the product of the lagging time and the rate of change of temperature.

If two bodies (e.g. a barometer and a thermometer) are immersed in the medium and λ_1 and λ_2 are their lagging times, their temperature difference will be $K(\lambda_1-\lambda_2)$.

§3. LAGGING TIME OF BODY OF CYLINDRICAL FORM IMMERSED IN LIQUID

A. *Theoretical computation*

The basis used for computing the lagging time theoretically is as follows :—

Consider an infinitely long cylindrical glass tube filled with mercury which is immersed in a medium whose temperature is rising at a constant rate. Let r_1 and r_2 be the internal and external radii of the glass tube; k_g, S_g and ρ_g the thermal conductivity, specific heat and density of glass; k_m, S_m and ρ_m the corresponding quantities for mercury; and let

$$\frac{k_g}{\rho_g S_g}=h_g^2,$$

$$\frac{k_m}{\rho_m S_m}=h_m^2.$$

After a certian lapse of time, the rate of flow of heat into the tube will become steady and the temperature at any point will rise at the same rate as the medium. The isothermal surfaces will be cylinders co-axial with the tube, and if θ is the temperature of an isothermal surface within the glass wall and of radius r, the differential equation expressing the flow of heat across this surface is

$$\frac{\partial\theta}{\partial t}=h_g^2\left(\frac{\partial^2\theta}{\partial r^2}+\frac{1}{r}\frac{\partial\theta}{\partial r}\right).$$

But $\frac{\partial\theta}{\partial t}$ is constant, so that for this condition

$$\frac{\partial^2\theta}{\partial r^2}+\frac{1}{r}\frac{\partial\theta}{\partial r}=C_g$$

where
$$C_g=\frac{\partial\theta}{\partial t}\bigg/h_g^2.$$

The integral of this equation is

$$\theta=\tfrac{1}{4}C_g r^2+A_g\log_e r+B_g,$$

in which A_g and B_g are constants*.

For an isothermal surface in the mercury, θ is given by the corresponding equation

$$\theta=\tfrac{1}{4}C_m r^2+A_m\log_e r+B_m,$$

in which

$$C_m=\frac{\partial\theta}{\partial t}\bigg/h_m^2\cdot \text{ and } A_m=0 \quad \text{since } \theta\neq\pm\infty \text{ when } r=0.$$

---

* Their values depend, of course, on the magnitude of $\partial\theta/\partial t$.

The equation giving the temperature of an isothermal surface in the mercury can also be obtained as follows. The heat which flows per unit time across an isothermal surface of radius r in the mercury column serves to raise the temperature of the mercury within this surface at the rate $\partial\theta/\partial t$.

Hence, $$2\pi r k_m \frac{\partial\theta}{\partial r} = \pi r^2 \rho_m S_m \frac{\partial\theta}{\partial t},$$

i.e. $$\frac{\partial\theta}{\partial r} = \tfrac{1}{2} r C_m,$$

or $$\theta = \tfrac{1}{4} C_m r^2 + B_m.$$

At any instant, whilst temperatures are changing at a steady rate, let

$$\theta = \theta_2 \quad \text{when} \quad r = r_2,$$

$$\theta = \theta_1 \quad \text{when} \quad r = r_1,$$

$$\theta = \theta_0 \quad \text{when} \quad r = 0,$$

and let λ_r be the lagging time in relation to the external wall of the tube of an isothermal surface of radius r within the mercury.

Then $$\lambda_r \frac{\partial\theta}{\partial t} = \theta_2 - \theta_r,$$

where $$\theta_r = \tfrac{1}{4} C_m r^2 + B_m,$$

or, since $$\theta_1 = \tfrac{1}{4} C_m r_1^2 + B_m,$$

$$\theta_r = \theta_1 - \tfrac{1}{4} C_m (r_1^2 - r^2).$$

Hence $$\lambda_r \frac{\partial\theta}{\partial t} = (\theta_2 - \theta_1) + (\theta_1 - \theta_r)$$

$$= \tfrac{1}{4} C_g (r_2^2 - r_1^2) + A_g \log_e \frac{r_2}{r_1} + \tfrac{1}{4} C_m (r_1^2 - r^2).$$

But $$\frac{\partial\theta}{\partial t} = C_g h_g^2 = C_m h_m^2,$$

and, therefore,

$$\lambda_r = \frac{1}{4h_g^2}(r_2^2 - r_1^2) + \frac{A_g}{C_g h_g^2} \log_e \frac{r_2}{r_1} + \frac{1}{4h_m^2}(r_1^2 - r^2).$$

The value of the constant A_g may be obtained by equating the heat flow per unit time across the inner wall of the glass tube to the heat required to raise the temperature of the mercury column at the constant rate $\partial\theta/\partial t$.

Thus, $$2\pi r_1 k_g \left(\frac{\partial\theta}{\partial r}\right)_{r=r_1} = \pi r_1^2 \rho_m S_m \frac{\partial\theta}{\partial t},$$

i.e. $$\left(\frac{\partial\theta}{\partial r}\right)_{r=r_1} = \tfrac{1}{2} r_1 \frac{\rho_m S_m}{\rho_g S_g} \cdot \frac{\rho_g S_g}{k_g} \cdot \frac{\partial\theta}{\partial t}$$

$$= \tfrac{1}{2} C_g \frac{\rho_m S_m}{\rho_g S_g} r_1.$$

Hence $$\tfrac{1}{2}C_g r_1 + \frac{A_g}{r_1} = \tfrac{1}{2}C_g \frac{\rho_m S_m}{\rho_g S_g} r_1,$$

which gives $$A_g = \tfrac{1}{2}C_g \left(\frac{\rho_m S_m - \rho_g S_g}{\rho_g S_g}\right) r_1^2.$$

Hence $$\lambda_r = \frac{1}{4h_g^2}(r_2^2 - r_1^2) + \frac{1}{2h_g^2}\left(\frac{\rho_m S_m - \rho_g S_g}{\rho_g S_g}\right) r_1^2 \log_e \frac{r_2}{r_1} + \frac{1}{4h_m^2}(r_1^2 - r^2)$$

or $$\lambda_r = \frac{\rho_g S_g}{4k_g}(r_2^2 - r_1^2) + \frac{\rho_m S_m - \rho_g S_g}{2k_g} r_1^2 \log_e \frac{r_2}{r_1} + \frac{\rho_m S_m}{4k_m}(r_1^2 - r^2). \quad \ldots\ldots(3)$$

This equation gives the lagging time, in relation to the external surface of the tube, of an isothermal surface of radius r within the mercury. Before theory can be compared with practice, it is necessary to determine the mean value of λ over the range $r = 0$ to $r = r_1$.

If λ_{r_1} is the lagging time of the inner wall of the glass tube in relation to the external surface, it is clear from the manner of derivation of λ_r (but see also § 4, B) that

$$\lambda_r = \lambda_{r_1} + \frac{\rho_m S_m}{4k_m}(r_1^2 - r^2),$$

so that the average lagging time λ_a is given by

$$\lambda_a = \lambda_{r_1} + \frac{\rho_m S_m}{4k_m} \cdot \frac{\int_{-r_1}^{+r_1} (r_1^2 - r^2)\,dr}{\int_{-r_1}^{+r_1} dr}$$

$$= \lambda_{r_1} + \frac{\rho_m S_m}{6k_m} r_1^2,$$

i.e.

$$\lambda_a = \frac{\rho_g S_g}{4k_g}(r_2^2 - r_1^2) + \frac{\rho_m S_m - \rho_g S_g}{2k_g} r_1^2 \log_e \frac{r_2}{r_1} + \frac{\rho_m S_m}{6k_m} r_1^2. \quad \ldots\ldots(4)$$

B. *Experimental determination*

The validity of this equation has been tested by determining experimentally the lagging time of a glass tube containing mercury and fitted with a thermometer. The bore of the tube used was 22·5 mm., the wall thickness 1·6 mm. and the length of the mercury column 815 mm. The thermometer passed through a small side tube sealed to the main tube at about the mid-point of the mercury column.

The tube was placed in a room whose temperature is maintained constant at about 35° C. (95° F.) and kept there until its temperature had become steady. Immediately outside the door of the room was placed a large iron vessel fitted with glass windows and filled with water at about 17° C. (62° F.). The lagging time was determined by transferring the tube as quickly as possible from the room to the water bath and taking correlated readings of temperature and time with the tube completely immersed in the water, commencing when the tube was lowered into the water and continuing until the temperature of the mercury had fallen to that of the bath.

Figure 1 shows the results obtained in graphical form. One graph is the plot against time of $\bar{\theta}$ and the other the plot against time of $\log_{10}(\bar{\theta}-\phi)$, $\bar{\theta}$ being the mercury temperature and ϕ the water temperature, which is effectively constant. (The bar notation is employed to indicate that the thermometer used gives the average temperature over the cross section of the mercury column). The value obtained for the lagging time was 25·3 seconds; a repetition of the experiment gave 25·5 seconds. When the values (in c.g.s. units) of r_1 and r_2 and of the physical constants k, S and ρ for glass and mercury are substituted in the equation given above for λ_a, the result obtained is 24 seconds.

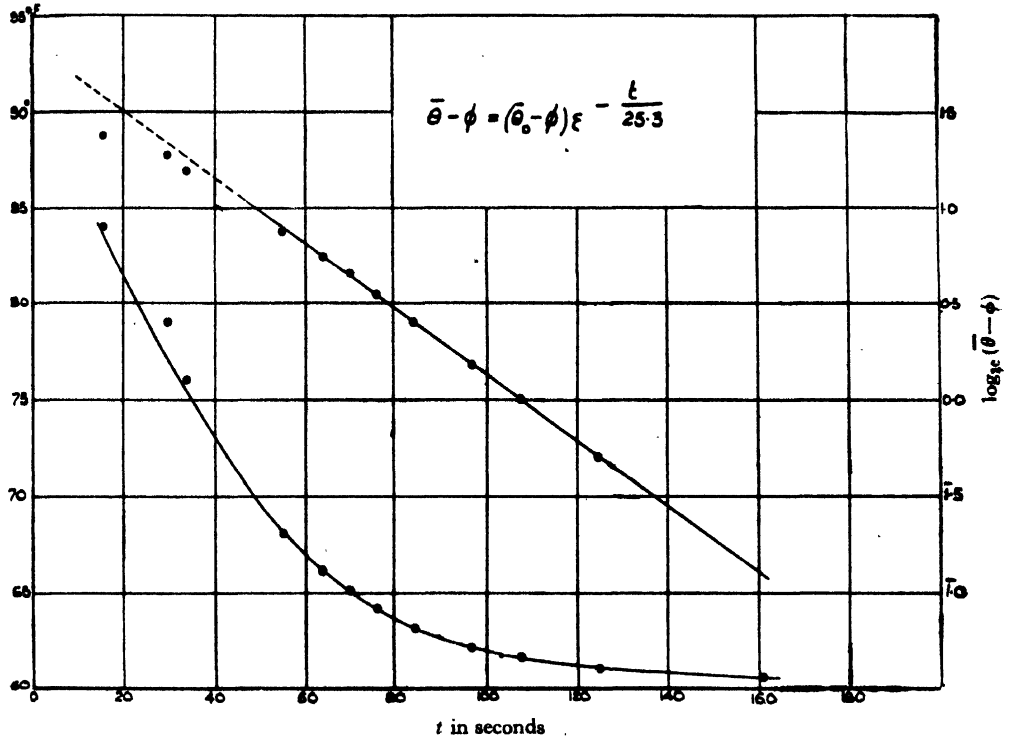


Figure 1. Observations on mercury column, in glass tube of bore 22·5 mm, immersed in water.

Attention may be drawn to the fact that the first three observational points plotted in the log graph fall distinctly below the straight line on which the other points lie. This is characteristic of all the curves which have been obtained in experiments on lagging time and is a reminder that the state to which the equations apply is not established instantly. In the experimental verification given above it was, of course, assumed that the outer wall of the glass tube was maintained throughout the observations at the temperature of the water bath; in point of fact, the rise of temperature of the bath during the experiment did not exceed 0°·1 C.

C. *Homogeneous cylinders*

For homogeneous cylinders the equation for λ is simpler. The temperature at a distance r from the axis is given (cf. §3, A) by

$$\theta=\tfrac{1}{4}Cr^2+B.$$

If θ_1 and θ_0 are the external and axial temperatures and R is the radius of the cylinder

$$\theta_1 - \theta_0 = \tfrac{1}{4}CR^2,$$

so that if λ_0 is the lagging time of the axis in relation to the surface of the cylinder,

$$\lambda_0 Ch^2 = \tfrac{1}{4}CR^2$$

or

$$\lambda_0 = \frac{1}{4h^2}R^2 = \frac{\rho S}{4k}R^2.$$

For a solid iron rod $\lambda_0 = 1{\cdot}7\,R^2$;

for a solid glass rod $\lambda_0 = 44R^2$;

and for a thin-walled glass tube containing toluene

$$\lambda_0 = 220R^2 \text{ approx.}$$

These equations give λ_0 in seconds when R is expressed in cm.

§4. LAGGING TIME IN STAGNANT AIR

A. *Theoretical computation*

The theory which has been given may be considered to deal with the lagging time of a body in a liquid medium. In order to extend it so that it applies to bodies immersed in a gaseous medium, it is necessary to introduce the law governing the rate of gain of heat by a body immersed in a gas at higher temperature.

Let

E = emissivity of the surface; θ = temperature of the surface;

A = area of the surface; ϕ = temperature of the medium.

Then the rate of gain of heat by the surface is $EA(\phi - \theta)$. The value of E depends, of course, on the state of the surface (e.g. whether highly polished or blackened) and on the degree of stagnation of the medium.

Consider a tube of liquid immersed in a gaseous medium whose temperature is rising at a steady rate and let

λ_A = lagging time of the axis in relation to the medium.

The lagging time of the axis in relation to the external surface of the tube will be identical with the lagging time in a liquid medium, e.g. water; it will, therefore, be referred to in what follows as the lagging time in water and will be denoted by λ_W. Then if, at any instant,

θ_A = temperature of the gaseous medium,

θ_2 = temperature of external surface of tube,

θ_0 = temperature at the axis, and

K = constant rate of change of temperature of the medium,

it follows that

$$\lambda_W \,.\, K = \theta_2 - \theta_0,$$

$$\lambda_A \,.\, K = \theta_A - \theta_0,$$

whence

$$\lambda_A = \lambda_W + \frac{\theta_A - \theta_2}{K}.$$

The heat gained per unit time by the external wall of the tube is used in raising the temperature of the tube and liquid by an amount K.

Using the same notation as before,

$$2\pi r_2 E(\theta_A - \theta_2) = \{\pi(r_2^2 - r_1^2)\rho_g S_g + \pi r_1^2 \rho_m S_m\} K,$$

so that

$$\frac{\theta_A - \theta_2}{K} = \frac{(r_2^2 - r_1^2)\rho_g S_g + r_1^2 \rho_m S_m}{2r_2 E}$$

Hence

$$\lambda_A = \lambda_W + \frac{(r_2^2 - r_1^2)\rho\, S_g + r_1^2 \rho_m S_m}{2r_2 E}, \qquad \ldots\ldots(5)$$

and is determined since λ_W is given by the earlier equations.

For homogeneous cylinders of radius R,

$$\lambda_A = \lambda_W + \frac{\rho S}{2E} R,$$

i.e.

$$\lambda_A = \frac{\rho S}{4k} R^2 + \frac{\rho S}{2E} R. \qquad \ldots\ldots(6)$$

Some interesting deductions may be made from equation (6). As examples, the lagging times in air of solid rods of glass and of iron will be compared with their lagging times in water. Let the radius of the rods be 1 cm.

The lagging times (to the axis) in water are:—

(1) for the glass rod 44 sec. (cf. § 3, C);
(2) for the iron rod 1·7 sec.

The lagging times in air at normal pressure are:—

(1) for the glass rod 1100 sec.;
(2) for the iron rod 2000 sec.

A value of 0·0002 c.g.s. units* has been adopted for E, the author's experience being that this value appears to be satisfactory for polished surfaces under ordinary atmospheric conditions.

It will be seen that, whereas the lagging time of the iron rod in water is negligible by comparison with that of the glass rod, the lagging time in air is almost twice as great. It is not until the radius of the rods exceeds 20 cm. that the better conductivity of the iron is able to offset the lower thermal capacity per unit volume of the glass. (This argument is of course subject to the assumption that $E = 0{\cdot}0002$ c.g.s. units in both cases.)

If the two 1 cm. rods are immersed in water whose temperature is rising at the rate of 1° C. per hour, the difference between the axial temperature and the temperature of the water is

0°·012 C. for the glass rod;
0°·0005 C. for the iron rod.

On the other hand, if the rods are immersed in air rising in temperature at the same rate, the lag of the axial temperature behind that of the air is

0°·3 C. for the glass rod;
0°·55 C. for the iron rod.

---

* Ingersoll and Zobel, *Mathematical Theory of Heat Conduction*, p. 163 (1913).

B. *Extension to more complex bodies of cylindrical form*

The computation of the lagging time of the barometric column of a barometer is more complex than any of the cases yet considered, since the glass tube is surrounded by a brass sheath. It can, however, be carried out by application of the principles already discussed, although it will simplify matters if two theorems of general application are first stated.

(a) Let 1, 2, 3 .... n be isothermal surfaces in a heterogeneous body immersed in a medium whose temperature is rising at the uniform rate K per unit time.

At any instant, let θ_r and θ_s be the temperatures of the isothermal surfaces r and s and let $\lambda_{r,s}$ be the lagging time of surface r behind surface s.

Then
$$\theta_2 - \theta_1 = \lambda_{1,2} \, . \, K,$$
$$\theta_3 - \theta = \lambda_{2,3} \, . \, K,$$

$$\theta_n - \theta_{n-1} = \lambda_{n-1,n} \, . \, K.$$
$$\therefore \quad \theta_n - \theta_1 = \{\lambda_{1,2} + \lambda_{2,3} + \dots + \lambda_{n-1,n}\} \, . \, K.$$

But
$$\theta_n - \theta_1 = \lambda_{1,n} \, . \, K,$$

so that
$$\lambda_{1,n} = \lambda_{1,2} + \lambda_{2,3} + \dots + \lambda_{n-1,n}. \qquad \dots\dots(7)$$

Particular examples of this theorem have already been given.

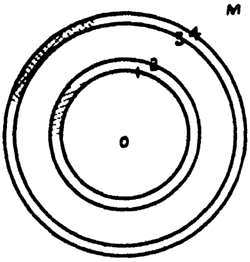


Figure 2. Section of barometer of normal design (not to scale).

(b) Consider a body bounded by a surface of area A in a gaseous medium whose temperature is rising at the rate K per unit time.

Let λ = lagging time of surface in relation to the medium;
ϕ = temperature of medium at any instant;
θ = temperature of surface at same instant;
U = thermal capacity of the body;
E = emissivity of the surface.

Then
$$\lambda \, . \, K = \phi - \theta$$

and
$$E \, . \, A \, .(\phi - \theta) = U.K.$$

Hence
$$\lambda = \frac{\phi - \theta}{K} = \frac{U}{EA}. \qquad \dots\dots(8)$$

Figure 2 shows a section of a barometer of normal design immersed in a gaseous medium M. O represents the axis of the barometer, 1 and 2 the surfaces of the glass tube and 3 and 4 the surfaces of the metal sheath. It is assumed that the glass and metal tubes are co-axial, so that 1, 2, 3 and 4 are isothermal surfaces.

Let r_1 = internal radius of the glass tube;
r_2 = external radius of the glass tube;
r_3 = internal radius of the metal tube;
r_4 = external radius of the metal tube.

Let $\lambda_{4,\mathrm{M}}$ be the lagging time of surface 4 in relation to the medium, $\lambda_{3,4}$ the lagging time of surface 3 in relation to surface 4, etc.

Then the lagging time of the axis behind the medium is given by

$$\lambda_{0,\mathrm{M}}=\lambda_{0,1}+\lambda_{1,2}+\lambda_{2,3}+\lambda_{3,4}+\lambda_{4,\mathrm{M}}.$$

But
$$\lambda_{0,1}+\lambda_{1,2}=\lambda_{0,2}=\lambda_{\mathrm{W}},$$
and has already been determined.

$$\lambda_{2,3}=\frac{U}{EA}=\frac{\rho_g S_g(r_2{}^2-r_1{}^2)+\rho_m S_m r_1{}^2}{2r_2E}.$$

$\lambda_{3,4}$ is obtained as follows:—

When the rate of change of temperature has become constant, temperatures in the metal sheath are given (cf. § 3, A) by the equation

$$\theta=\tfrac{1}{4}C_b r^2+A_b\log_e r+B_b,$$

in which the suffix b denotes that the constants are for brass.

The flow of heat per unit time across unit length of surface 3 is

$$2\pi r_3 k_b\left(\frac{\partial\theta}{\partial r}\right)_{r=r_3}=2\pi r_3 k_b\left(\tfrac{1}{2}C_b r_3+\frac{A_b}{r_3}\right).$$

The thermal capacity of the medium between surfaces 2 and 3 is negligibly small and it may be assumed that the heat flow from surface 3 is entirely absorbed by the barometer tube.

Hence
$$2\pi r_3 k_b\left(\tfrac{1}{2}C_b r_3+\frac{A_b}{r_3}\right)=2\pi r_2 E(\theta_3-\theta_2),$$

where θ_2 and θ_3 are simultaneous values of the temperatures of surfaces 2 and 3.

But
$$\theta_3-\theta_2=\lambda_{2,3}\cdot\frac{\partial\theta}{\partial t}=\lambda_{2,3}C_b h_b{}^2,$$

so that
$$\tfrac{1}{2}r_3{}^2 k_b C_b+k_b A_b=r_2E\lambda_{2,3}C_b h_b{}^2,$$

i.e.
$$A_b=\tfrac{1}{2}C_b\left(\frac{2r_2E}{k_b}\lambda_{2,3}h_b{}^2-r_3{}^2\right).$$

Hence
$$\lambda_{3,4}=\frac{\theta_4-\theta_3}{C_b h_b{}^2}=\frac{C_b(r_4{}^2-r_3{}^2)}{4C_b k_b{}^2}+\frac{C_b\left(\frac{2r_2E}{k_b}\lambda_{2,3}h_b{}^2-r_3{}^2\right)}{2C_b h_b{}^2}\log_e\frac{r_4}{r_3},$$

i.e.
$$\lambda_{3,4}=\frac{\rho_b S_b(r_4{}^2-r_3{}^2)}{4k_b}-\frac{\rho_b S_b r_3{}^2}{2k_b}\log_e\frac{r_4}{r_3}+\frac{r_2E}{k_b}\log_e\frac{r_4}{r_3}\lambda_{2,3}.$$

It is unnecessary to substitute for $\lambda_{2,3}$ since, for purposes of calculation, $\lambda_{2,3}$ must be evaluated and the value obtained can be introduced when computing $\lambda_{3,4}$.

$$\lambda_{4,\mathrm{M}}=\frac{U}{EA}=\frac{\rho_b S_b(r_4{}^2-r_3{}^2)+\rho_g S_g(r_2{}^2-r_1{}^2)+\rho_m S_m r_1{}^2}{2r_4E}.$$

This may be simplified for computing purposes as follows:—

$$\lambda_{4,\mathrm{M}}=\frac{\rho_b S_b(r_4^2-r_3^2)}{2r_4E}+\frac{\rho_g S_g(r_2^2-r_1^2)+\rho_m S_m r_1^2}{2r_2E}\times\frac{r_2}{r_4}$$

$$=\frac{\rho_b S_b(r_4^2-r_3^2)}{2r_4E}+\frac{r_2}{r_4}\lambda_{2,3}.$$

Hence, finally,

$$\begin{aligned}\lambda_{0,\mathrm{M}}=&\frac{\rho_g S_g(r_2^2-r_1^2)}{4k_g}+\frac{\rho_m S_m-\rho_g S_g}{2k_g}r_1^2\log_e\frac{r_2}{r_1}+\frac{\rho_m S_m r_1^2}{4k_m}\\&+\frac{\rho_g S_g(r_2^2-r_1^2)+\rho_m S_m r_1^2}{2r_2E}\\&+\frac{\rho_b S_b(r_4^2-r_3^2)}{4k_b}-\frac{\rho_b S_b r_3^2}{2k_b}\log_e\frac{r_4}{r_3}+\frac{r_2E}{k_b}\log_e\frac{r_4}{r_3}\lambda_{2,3}\\&+\frac{\rho_b S_b(r_4^2-r_3^2)}{2r_4E}+\frac{r_2}{r_4}\lambda_{2,3}.\end{aligned}\qquad\ldots\ldots(9)$$

It remains to point out that in order to obtain a mean lagging time for the mercury column instead of the lagging time to the axis, the denominator of the last term in the expression for λ_W is changed from $4k_m$ to $6k_m$ (cf. end of § 3, A).

For purposes of calculation, the form given above for $\lambda_{0,\mathrm{M}}$ is most suitable; it is of interest, however, to express $\lambda_{0,\mathrm{M}}$ in terms collected separately for the mercury, glass and metal components, viz.,

$$\begin{aligned}\lambda_{0,\mathrm{M}}=&\rho_m S_m r_1^2\left\{\frac{1}{4k_m}+\frac{1}{2k_g}\log_e\frac{r_2}{r_1}+\frac{1}{2r_2E}+\frac{1}{2k_b}\log_e\frac{r_4}{r_3}+\frac{1}{2r_4E}\right\}\\&+\rho_g S_g(r_2^2-r_1^2)\left\{\frac{1}{4k_g}+\frac{1}{2r_2E}+\frac{1}{2k_b}\log_e\frac{r_4}{r_3}+\frac{1}{2r_4E}\right\}-\frac{\rho_g S_g r_1^2}{2k_g}\log_e\frac{r_2}{r_1}\\&+\rho_b S_b(r_4^2-r_3^2)\left\{\frac{1}{4k_b}+\frac{1}{2r_4E}\right\}-\frac{\rho_b S_b r_3^2}{2k_b}\log_e\frac{r_4}{r_3}.\end{aligned}\qquad\ldots\ldots(10)$$

The following table gives the lagging times of the barometric columns of three barometers of different sizes, but all of ordinary design, as determined by means of equation (9). The first barometer is one of Meteorological Office pattern, the other two are working standard barometers used at the National Physical Laboratory:—

Table 1

Dimensions of barometer	Lagging time in air (minutes)
Internal diameter of glass tube = 8 mm. External diameter of brass tube = 25 mm.	20
Internal diameter of glass tube = 16 mm. External diameter of brass tube = 34 mm.	35
Internal diameter of glass tube = 20 mm. External diameter of brass tube = 39 mm.	45

A value of 0·0002 c.g.s. units for *E* has again been adopted in making the calculations. It should be pointed out that the brass tube of a barometer of ordinary design is slotted at its upper end and that no allowance for this departure from the theoretical assumptions has been made in the calculations.

C. *Experimental determinations*

Experimental determinations of the lagging times of mercury barometers have given results in reasonably good agreement with the values tabulated above, but the measurements are difficult to make and the best confirmation of the principles discussed has been the substantial improvement in accuracy obtained when the temperature of the barometer is determined by means of an appropriately lagged thermometer, i.e. one which has a lagging time equal to that of the barometer as obtained by calculation. A lagging of calculable effect can be provided by immersing the bulb of the thermometer in a glass tube containing mercury and surrounding the tube with one or more metal sheaths. The lagging time is computed in the manner already described and the following table gives results for typical cases:—

Table 2

Description of lagging	Lagging time in air (minutes)
Unshielded thermometer with thin-walled glass bulb of diameter 5 mm.*	3
Same thermometer shielded with brass sheath of diameter 20 mm., wall thickness 1 mm.	10
Bulb of thermometer immersed in mercury contained in glass tube of bore 16 mm. and wall thickness 2·5 mm.	20
Thermometer in same tube with iron sheath of internal diameter 30 mm. and wall thickness 4 mm.	52
Thermometer in same tube with iron sheath as above and an outer brass sheath of internal diameter 44 mm. and wall thickness 4 mm.	95

* The lagging time of this thermometer in water is 3 to 4 seconds.

These values also are based on a value of 0·0002 c.g.s. units for *E*.

It is of interest to show the contributions of the various terms λ_W, $\lambda_{2,3}$, etc. to the total lagging time $\lambda_{0,M}$. For the thermometer with the single iron sheath cited in the table above,

$$\lambda_W = 30\,\text{sec.} \qquad \lambda_{2,3} = 1100\,\text{sec.}$$

$$\lambda_{3,4} \text{ is quite negligible} \quad \text{and} \quad \lambda_{4,M} = 2000\,\text{sec.}$$

The lagging time between the surfaces of the metal sheath can then, for all practical purposes, be neglected and this leads to a considerable simplification of the calculations for cases such as the last referred to in table 2. The lagging time for this thermometer is

$$\lambda_{0,M}=\lambda_{0,1}+\lambda_{1,2}+\lambda_{2,3}+\lambda_{3,4}+\lambda_{4,5}+\lambda_{5,6}+\lambda_{6,M},$$

$$\lambda_{0,1}+\lambda_{1,2}=\lambda_W;$$

$\lambda_{3,4}$ and $\lambda_{5,6}$ can be neglected.

By analogy with equation (9) the expressions for the remaining terms are

$$\lambda_{2,3}=\frac{\rho_g S_g(r_2^2-r_1^2)+\rho_m S_m r_1^2}{2r_2 E},$$

$$\lambda_{4,5}=\frac{\rho_f S_f(r_4^2-r_3^2)}{2r_4 E}+\frac{r_2}{r_4}\lambda_{2,3},$$

$$\lambda_{6,M}=\frac{\rho_b S_b(r_6^2-r_5^2)}{2r_6 E}+\frac{r_4}{r_6}\lambda_{4,5}.$$

Practical determinations of the lagging times in air of thermometers immersed in mercury and surrounded by one or two metal sheaths (dimensions of glass and metal tubes as given in table 2) have been made. The experimental procedure was similar to that already described in relation to the mercury tube. The iron chamber was used again, but this time with a view to obtaining reasonably stagnant air conditions. The thermometers (one with a single sheath, the other with two) were mounted on a slab of cork 2 inches thick, the stems of the thermometers being 3 inches apart. Annular grooves of small depth were cut in the cork to locate the mercury tubes and the metal sheaths, a little plasticene being used to provide grip. The cork slab was suspended by means of brass strips from a wooden support shaped sc as to form a lid for the iron chamber. For purposes of comparison an unshielded thermometer was suspended between the other two.

Experiments were made under both falling and rising conditions of temperature. Several hours before commencing the test under falling conditions, the thermometers were placed in a room maintained at 35° c. approximately and the iron chamber was placed immediately outside the door of the room. For the other experiment, the iron chamber was placed in the hot room whilst the thermometers were kept in an adjoining room which was maintained at 20° c.

The actual thermometric readings obtained in the experiments are not given here but the logarithmic plots are shown in figures 3 and 4. The results obtained were:—

Thermometer with one sheath	
Falling temperature	49 minutes
Rising temperature	50 minutes
Thermometer with two sheaths	
Falling temperature	77 minutes
Rising temperature	78 minutes

The lagging time of the unshielded thermometer was, very closely, 3 minutes. The calculated values for the two shielded thermometers are as given in table 2, viz. 52 and 95 minutes. It must be mentioned that the sheaths and mercury tubes

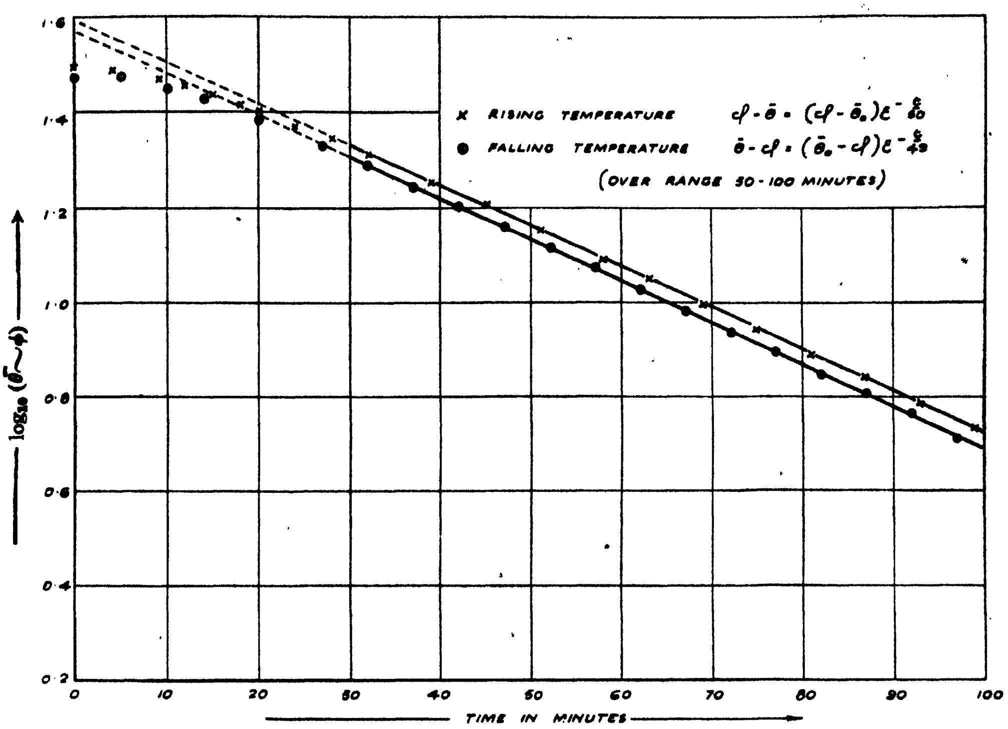


Figure 3. Observations on mercury-immersed thermometer, with single metal sheath, in stagnant air.

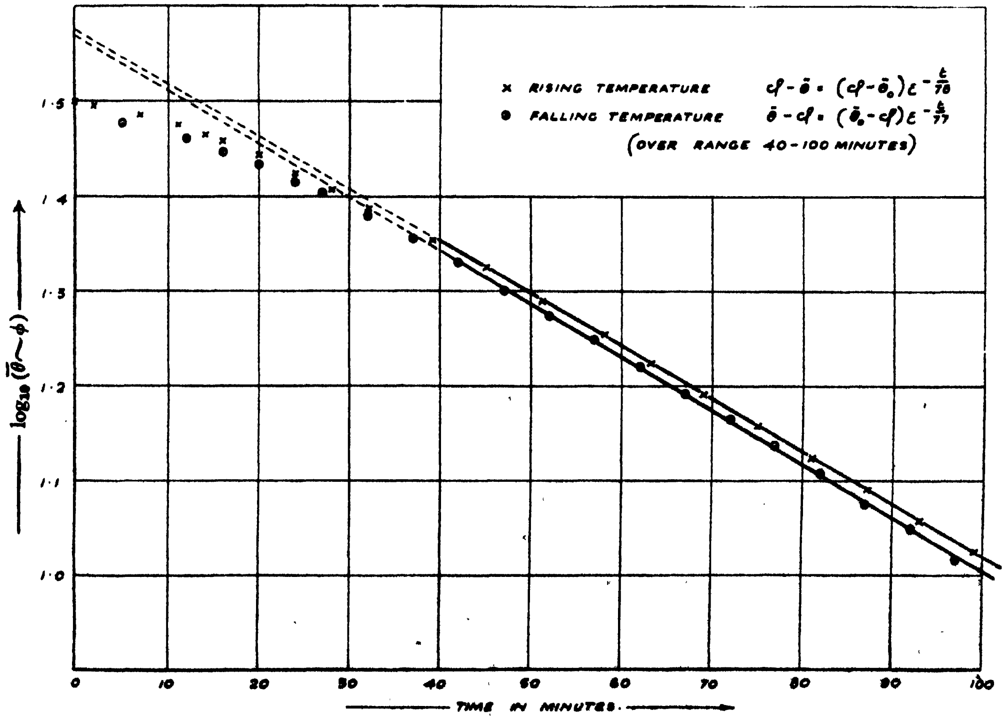


Figure 4. Observations on mercury-immersed thermometer, with double metal sheath, in stagnant air.

used were only about 4 inches long, so that the end effects, particularly for the thermometer with two sheaths, were by no means negligible. The agreement between theory and practice is, however, reasonably good, bearing in mind the uncertainty attaching to the value of E.

§ 5. ACKNOWLEDGMENTS

The work described above was carried out as part of the research programme of the National Physical Laboratory, and this paper is published by permission of the Director of the Laboratory. The author desires to acknowledge the comments and suggestions he received from Mr. F. A. Gould and the assistance in the experimental work rendered by Mr. F. D. Jones.

DISLOCATIONS IN A SIMPLE CUBIC LATTICE

By F. R. N. NABARRO,
Royal Society Warren Research Fellow

Communicated by N. F. Mott, F.R.S. ; *MS. received* 23 *September* 1946

ABSTRACT. The properties of dislocations are calculated by an approximate method due to Peierls. The width of a dislocation is small, displacements comparable with the interatomic distance being confined to a few atoms. The shear stress required to move a dislocation in an otherwise perfect lattice is of the order of a thousandth of the " theoretical " shear strength. The energy and effective mass of a single dislocation increase logarithmically with the size of the specimen. A pair of dislocations of opposite sign in the same glide plane cannot be in stable equilibrium unless they are separated by a distance of the order of 10 000 lattice spacings. If an external shear stress is applied there is a critical separation of the pair of dislocations at which they are in unstable equilibrium. The energy of this unstable state is the activation energy for the formation of a pair of dislocations. It depends on the external shear, and for practical stresses is of the order of 7 electron volts per atomic plane.

The size and energy of dislocations in real crystals are unlikely to differ greatly from those calculated : the stress required to move a dislocation and the critical separation of two dislocations may be seriously in error.

§ 1. INTRODUCTION.

VOLTERRA'S theory of elastic dislocations has been made the basis of a theory of the plastic deformation of crystals. This theory has explained many of the characteristics of plastic deformation, but Volterra's classical assumptions of a continuous medium obeying Hooke's Law even under large strains do not allow a detailed investigation of the displacement of the atoms in the core of a dislocation in a real crystal. Peierls (1940) has shown that by using suitable approximations it is possible to take account of the periodic structure of the crystal in the glide plane, and to determine the size of a dislocation in a simple cubic lattice and the shear stress necessary to move such a dislocation across its glide plane. The first part of this paper amplifies the previous arguments and corrects certain errors in Peierls's paper. In the second part, the solution for a single dislocation is extended to the case of a pair of dislocations of opposite sign held in equilibrium in the same glide plane by an external shear stress. The energy of this system, which is the activation energy for the formation of a pair of dislocations in a stressed crystal, is determined. Finally, a brief discussion is given of the relation of the properties of this idealized model to those of a real crystal.

§ 2. THE MODEL

The model considered is a simple cubic lattice containing a dislocation of the kind which Burgers (1940) calls "a dislocation of edge type". It is shown in figure 1. The slip plane $P(z=0)$ divides the crystal into an upper part a and a lower part b. These are symmetrical about the vertical plane $S(x=0)$. The central plane S lies in a lattice plane in the upper half-crystal a, and half way between two lattice planes in b.

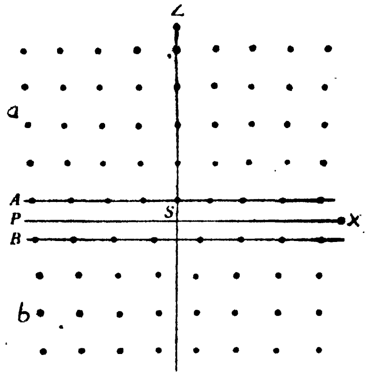


Figure 1. A single dislocation in an unstrained crystal.

The y axis at right angles to the plane of the figure (i.e. the line in which the plane S meets the slip plane P) is called the *dislocation line*. In the neighbourhood of this line the atoms of a are moved inwards and those of b are moved apart, so that at great distances from the dislocation line the two planes A and B, which are immediately above and below the slip plane, are again in correct alignment, A containing, however, one row of atoms more than B. This type of dislocation, which may be formed by cutting a perfect crystal along the half-plane $x=0$, $z>0$, and introducing a single extra layer of atoms into this half-plane, can also be represented as a Taylor dislocation, formed by cutting the same crystal along the half-plane $z=0$, $x<0$ and sliding the upper layer of atoms A to the right over the lower layer B until the relative displacement of the planes at large distances from the dislocation line is one lattice spacing (Taylor, 1934). In Burgers' form, the crystal containing a dislocation is in equilibrium with no external forces, whereas external forces are required to maintain the equilibrium of a crystal containing a dislocation of Taylor's form.

The atoms in the plane A are then subject to two forces : (i) the interaction with other atoms in the half-crystal a; this force tends to spread the compression uniformly over the plane A, i.e. to extend the dislocation: (ii) the interaction with atoms in b, particularly with those in the adjacent layer B; this has a tendency to bring as many atoms of A as possible into correct alignment with B, i.e. to shorten the dislocation. In equilibrium these two forces balance.

If the extension of the dislocation is large compared with the atomic distance d, both the horizontal displacement u and the vertical displacement w of the atoms in A vary slowly from atom to atom. The relative displacement of neighbouring atoms within each half-crystal is then much smaller than d. In these circumstances each half-crystal may be considered as an elastic continuum. Moreover, for the sake of simplicity, it will be assumed to be elastically isotropic. Then the force (i) is simply the force that has to be applied to the plane surface of an elastic continuum in order to make its horizontal displacement at the surface equal to $u(x)$. This force can be obtained by the usual methods of the theory of elasticity.

To the same degree of approximation, it may be assumed that the horizontal component of the force (ii) depends only on the horizontal displacement of the atoms of A relative to the atoms of B immediately underneath. If $u(x)$ and $\bar{u}(x)$ are the two displacements, the force acting on a surface element near x is a periodic function of $u-\bar{u}$ with the period d. In a first approximation it may be represented as a simple sinusoidal function of the form

$$\text{const.} \sin 2\pi(u-\bar{u})/d. \qquad \ldots\ldots(1)$$

The probable deviation of the law of force from this form, and the consequences of such a deviation, are discussed in §7. If the simple sine law is a sufficient approximation, the constant can be found from the shear modulus, provided it is assumed that the force arising between two lattice planes in a state of shear is independent of the displacement of the other lattice planes. This assumption is not strictly correct, but it is probably a reasonable approximation. By considering a small shear of angle $(u-\bar{u})/d$ it may be seen that the constant is $-\mu/2\pi$.

The dislocation is considered to be very long in the direction of the y axis, and the problem is accordingly one of plane strain. Further, it is assumed that the vertical component of the force (ii) depends only on the relative vertical displacement in A with respect to that in B. Since the tangential forces applied to A and B are equal and opposite, their vertical displacements are equal, and the vertical force (ii) vanishes. This assumption again is not actually correct, since in places where the atoms in the two planes are out of alignment, the equilibrium value of their vertical distance will obviously be changed. A separate discussion is required in order to take this effect into account, and it depends on the detailed structure of the crystal lattice.

§3. THE GOVERNING EQUATION

It is first necessary to find the integral equation connecting the displacement $u(x')$ of a point x' on the surface A with the tangential stress $p_{zx}(x)$ applied to this surface, assuming that the strains e_{xy}, e_{yy}, e_{zy} vanish across the surface A in accordance with the assumption of §2.

The problem may be solved by assuming that the stresses are expressed in terms of a stress function χ in the usual form:

$$\left.\begin{aligned} p_{xx} &= \frac{\partial^2\chi}{\partial z^2}, \\ p_{xz} &= -\frac{\partial^2\chi}{\partial x\partial z}, \\ p_{zz} &= \frac{\partial^2\chi}{\partial x^2}, \end{aligned}\right\} \quad \ldots\ldots\ldots(2)$$

while for plane strain the components of strain can be expressed in terms of the components of stress in the form

$$4\mu(\lambda+\mu)e_{xx} = (\lambda+2\mu)p_{xx} - \lambda p_{zz}, \text{ etc.,}$$

where λ and μ are Lamé's elastic contants. Expressing λ in terms of μ and Poisson's ratio σ by means of the relation

$$\frac{\lambda}{\mu} = \frac{2\sigma}{1-2\sigma}$$

leads to

$$2\mu e_{xx} = (1-\sigma)p_{xx} - \sigma p_{zz}, \quad \ldots\ldots(3)$$

The integral equation may be obtained by choosing a stress function

$$\chi = Z(e^{-mZ}\cos mx - 1), \quad \ldots\ldots(4)$$

where $Z = z - \frac{1}{2}d$.

This stress function is easily shown to satisfy the equation $\nabla^4\chi = 0$, and leads to

$$\begin{aligned} p_{xx} &= -m(2-mZ)e^{-mZ}\cos mx, \\ p_{xz} &= m(1-mZ)e^{-mZ}\sin mx, \\ p_{zz} &= -m^2Ze^{-mZ}\cos mx. \end{aligned}$$

On the surface A, $Z=0$, and p_{zz} vanishes across this surface, as has been assumed. The tangential stress is given by

$$p_{xz} = m\sin mx, \quad \ldots\ldots(5)$$

while $p_{xx} = -2m\cos mx.$

It follows from (3) that

$$\mu e_{xx} = -(1-\sigma)m\cos mx$$

and by integration that

$$\mu u(x) = -(1-\sigma)\sin mx. \quad \ldots\ldots(6)$$

A more general displacement $u(x)$ may now be expressed as a Fourier integral of the form

$$u(x) = \frac{1}{\pi}\int_0^\infty\int_{-\infty}^\infty \cos m(x'-x)\,.\,u(x')dx'dm. \quad \ldots\ldots(7)$$

Since the equations of elasticity are linear, the components of displacement corresponding to each value of m may be considered separately. A typical component is of the form

$$\frac{1}{\pi}\int_{-\infty}^{\infty} \cos m(x'-x) \,.\, u(x')dx'$$

$$= \frac{1}{\pi}\int_{-\infty}^{\infty} \cos mx' \,.\, u(x')dx' \,.\, \cos mx$$

$$+ \frac{1}{\pi}\int_{-\infty}^{\infty} \sin mx' \,.\, u(x')dx' \,.\, \sin mx.$$

Each term is of the same form as (6) or the corresponding expression involving $\cos mx$, and the corresponding tangential stresses are, therefore, given by (5) and the corresponding expression. The resultant tangential stress is given formally by

$$p_{xz} = -\frac{\mu}{\pi(1-\sigma)}\int_0^{\infty}\int_{-\infty}^{\infty} m \cos m(x'-x) \,.\, u(x')dx'dm. \qquad \ldots\ldots(8)$$

Integrating by parts, and presupposing that $u(\infty) = u(-\infty) = 0$, gives

$$\pi(1-\sigma)p_{xz}/\mu = \int_0^{\infty}\int_{-\infty}^{\infty} \sin m(x'-x)\frac{du}{dx'}dx'dm.$$

A second formal integration by parts leads to

$$\pi(1-\sigma)p_{xz}/\mu = \lim_{m\to\infty}\int_{-\infty}^{\infty} \frac{1-\cos m(x'-x)}{x'-x}\frac{du}{dx'}dx', \qquad \ldots\ldots(9)$$

and if du/dx' is a sufficiently regular function, the term involving $\cos m(x'-x)$ tends to 0 as m tends to ∞. (For a rigorous discussion of this type of transformation see Titchmarsh, 1937.) The general relation now becomes

$$p_{xz} = \frac{\mu}{\pi(1-\sigma)}\int_{-\infty}^{\infty} \frac{1}{x'-x}\frac{du}{dx'}dx', \qquad \ldots\ldots(10)$$

where the Cauchy principal value of the integral is taken.

This equation may be compared with that derived from equation (1), which is

$$p_{xz} = -\frac{\mu}{2\pi}\sin\left[2\pi(u-\bar{u})/d\right].$$

Writing $\phi = u - \bar{u} = 2u$, elimination of p_{xz} yields

$$\int_{-\infty}^{\infty} \frac{1}{x-x'}\frac{d\phi}{dx'}dx' = (1-\sigma)\sin\frac{2\pi\phi}{d}. \qquad \ldots\ldots(11)$$

This equation differs slightly from that previously given.

§4. THE SOLUTION FOR A SINGLE DISLOCATION

The solution which represents a single dislocation with its centre at the origin as in figure 1, is

$$\frac{\phi}{d} = -\frac{1}{\pi}\tan^{-1}\frac{2x(1-\sigma)}{d}. \qquad \ldots\ldots(12)$$

This solution was given by Peierls. It may be verified by substituting in (11), resolving the integral into partial fractions, and taking the Cauchy principal value of the divergent term.

The width of the dislocation is small, the displacement falling to half its extreme value at a distance $d/2(1-\sigma)$ from the central plane S. The original assumption that the dislocation is spread over a large number of atoms has thus led to a contradiction and, if the sinusoidal law of force (1) is valid and the neglect of vertical forces is justified, a dislocation must be confined to a region of linear dimensions only a few atomic radii.

The components of displacement in parts of the crystal away from the slip plane may be obtained by expressing (12) as a Fourier integral in the form

$$u = \tfrac{1}{2}\phi = -\frac{d}{2\pi}\int_0^\infty e^{-\frac{md}{2(1-\sigma)}} \frac{\sin mx}{m}\,dm. \qquad \ldots\ldots(13)$$

Comparing (13) with (6) and (4) shows that the stress function is

$$\chi = -\frac{\mu d}{2\pi(1-\sigma)}\int_0^\infty e^{-\frac{md}{2(1-\sigma)}} Z\,\frac{e^{-mZ}\cos mx - 1}{m}\,dm. \qquad \ldots\ldots(14)$$

The displacements corresponding to (4) are given (Love, 1944) by

$$\left.\begin{aligned} 2\mu u &= [mZ-2(1-\sigma)]e^{-mZ}\sin mx,\\ 2\mu w &= [(mZ-1)+2(1-\sigma)]e^{-mZ}\cos mx + 1 - 2(1-\sigma), \end{aligned}\right\} \qquad \ldots\ldots(15)$$

and the displacements corresponding to (14) follow immediately. The singularity at the origin, which is represented in elastic dislocation theory by a factor $[x^2+z^2]$ in the denominator, now disappears, and this factor is replaced by $[x^2+(z+\zeta)^2]$, where $\zeta = d/2(1-\sigma)$.

The energy of unit length of such a dislocation in an infinite crystal is infinite, for the shearing strain at a large distance r from the axis is of the form B/r. The energy density is of the order $1/r^2$, and the total elastic energy of order $\int^\infty (1/r^2)2\pi r\,dr$, which diverges. A finite energy is obtained only by considering a pair of dislocations of opposite sign.

A dislocation has an effective mass, for as it moves across the glide plane the displacements of all points in the body alter. If the dislocation moves across the glide plane with velocity V, the components of velocity of any other point in the body are $V\,\partial u/\partial x$, $V\,\partial w/\partial x$. If the density of the crystal is ρ, the kinetic energy associated with unit length of dislocation is

$$\tfrac{1}{2}\rho V^2 \iint \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial x}\right)^2\right] dx\,dz.$$

This is of the same form as the elastic energy, and diverges logarithmically.

If the dislocation is situated in a finite crystal of linear dimensions L, the elastic energy of unit length of the dislocation is of order $\mu d^2 \log L/d$, and its effective mass per unit length is of order $\rho d^2 \log L/d$.

§5. THE SOLUTION FOR A PAIR OF DISLOCATIONS

It has been suggested that plastic flow may be initiated by the simultaneous production of a pair of dislocations of opposite sign moving in the same glide plane. The form of a crystal under external shear stress, and containing such a pair of dislocations, is shown in figure 2. It is similar to the ***Verhakung*** discussed by Dehlinger (1929). Such a system cannot be in stable equilibrium. The external stress causes the dislocations to separate, and their separation leads to a shearing motion of the upper and lower parts of the crystal which yields to the external stress. The stress field surrounding each dislocation opposes the external stress in the neighbourhood of the other dislocation, and gives rise to an attraction between the dislocations. For a given applied stress T there is a critical separation $r(T)$ at which the dislocations are in unstable equilibrium, and it is this state of unstable equilibrium which is calculated.

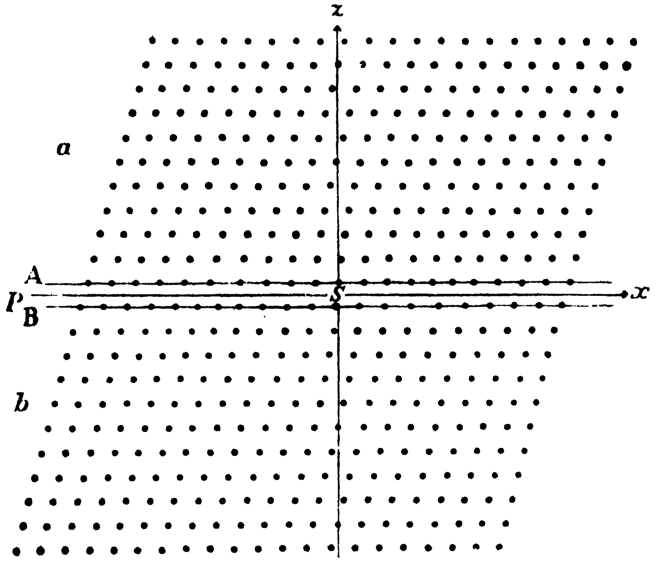


Figure 2. A pair of dislocations in a sheared crystal. The angle of shear shown is ten times as great as that required to maintain equilibrium in the planes A and B.

In the case of figure 1 it was convenient to measure u and $\bar{u}$ from positions such that $\phi=0$ represented the greatest possible misfit of the lattices above and below the glide plane: for the double dislocation of figure 2 it is more convenient to represent this misfit by $\phi=\frac{1}{2}d$ and exact registration of the lattices by $\phi=0$ or $\phi=d$. This changes the sign of the sine function in equations (1) and (11).

The solution for two dislocations is then

$$\frac{\phi}{d}=\frac{1}{\pi}\cot^{-1}\left[\frac{(1-\sigma)^2\sin\theta}{d^2}x^2-\cot\theta\right]+\frac{\theta}{4\pi}, \qquad \text{......(16)}$$

where θ is a parameter defining the separation of the dislocations and that value of $\cot^{-1}$ is taken which lies between 0 and π. This solution may be verified by substitution.

The displacements in rows A and B in figure 2 correspond to $\theta = \pi/10$. Substituting in (10), the tangential stress required to maintain this displacement is

$$p_{xz} = \frac{\mu}{2\pi}\left(\sin\frac{2\pi\phi}{d} - \sin\frac{\theta}{2}\right), \qquad \ldots\ldots(17)$$

whereas the force produced by the atoms in the other half of the crystal is

$$p_{xz} = \frac{\mu}{2\pi}\sin\frac{2\pi\phi}{d}. \qquad \ldots\ldots(18)$$

The actual tangential stress (18) exceeds the stress (17) required to maintain the displacement (16) by a constant quantity $(\mu/2\pi)\sin\frac{1}{2}\theta$. The upper and lower halves of the crystal can only remain in equilibrium if external stresses of this amount are applied to them. If θ is small, this stress, $\mu\theta/4\pi$, produces a uniform shearing strain of $\theta/4\pi$, which agrees with the strain between rows A and B given by (16) for large values of x.

If θ is small the two dislocations are widely separated. Their centres may be defined as the points $x = \pm d(\cos\theta)^{\frac{1}{2}}/(1-\sigma)\sin\theta$, or for small θ, $x = \pm d/(1-\sigma)\theta$.

It is now possible to calculate the energy of this pair of dislocations. Since the dislocations are embedded in an infinite crystal under a uniform shear stress $\mu\theta/4\pi$, the total energy of the system is infinite. The energy which it is important to calculate is the activation energy required to produce the equilibrium configuration of figure 2 when the crystal is already uniformly stressed. Once this energy is supplied, further separation of the dislocations releases energy.

The energy of the crystal in figure 2 differs from the energy of the uniformly strained crystal by three contributions:

(a) the forces acting across the boundary P do work on each half-crystal;

(b) the potential energy of attraction between the atoms in rows A and B is increased;

(c) the separation of the dislocations represents a shear of one half-crystal over the other, and in this motion work is done by the external shearing forces.

These contributions may be evaluated separately.

(a) *Work done by forces acting across P*

Each half of the crystal is assumed to obey Hooke's Law, and its elastic energy may be expressed in terms of the work done by the forces acting across the plane P by integrals of the form $\frac{1}{2}\int u_x p_{xz}\,dx$. If the energy is referred to the crystal under uniform shear as zero, the total energy is

$$-2\int_{-\infty}^{\infty}\tfrac{1}{2}[u(x)p_{xz}(x) - u(\infty)p_{xz}(\infty)]\,dx$$

$$= -\frac{\mu}{4\pi}\int_{-\infty}^{\infty}\left[\phi\sin\frac{2\pi\phi}{d} - \frac{d\theta}{4\pi}\sin\frac{\theta}{2}\right]dx, \qquad \ldots\ldots(19)$$

where ϕ is given by (16).

The expression (19) may be evaluated by writing $x = d(\cos\theta)^{\frac{1}{2}}y/(1-\sigma)\sin\theta$, and becomes

$$-\frac{\mu d^2(\cos\theta)^{\frac{1}{2}}}{4\pi^2(1-\sigma)\sin\theta}\int_{-\infty}^{\infty}[\chi\sin 2\chi-\tfrac{1}{4}\theta\sin\tfrac{1}{2}\theta]\,dy, \qquad \ldots\ldots 20)$$

where
$$\cot(\chi-\tfrac{1}{4}\theta)=(y^2-1)\cot\theta.$$

It is easily shown that the integral in (20) is equal to the integral of

$$\frac{1}{2i}\cdot\frac{2(y^2-1)\cot\theta\cos\frac{1}{2}\theta+[(y^2-1)^2\cot^2\theta-1]\sin\frac{1}{2}\theta}{(y^2-1)^2\cot^2\theta+1}\log\frac{(y^2-1)\cot\theta+i}{(y^2-1)\cot\theta-i}$$
$$+\frac{\theta}{2}\cdot\frac{(y^2-1)\cot\theta\cos\frac{1}{2}\theta-\sin\frac{1}{2}\theta}{(y^2-1)^2\cot^2\theta+1}$$

taken anti-clockwise round the contour of figure 3. This function has essential singularities of the form $(\log z)/z$ at the points P, $Q=\pm(\cos\theta)^{-\frac{1}{2}}e^{\pm\frac{1}{2}i\theta}$. If cuts are made joining P and Q to the point $i\eta$, where η is real and very small, and corresponding cuts in the lower half-plane, the function is single-valued and the logarithm changes by $\pm i\pi$ on crossing the boundaries OP, OQ. The term independent of the logarithm is always single-valued, and has only simple poles at P and Q. The residues are equal and opposite, and this term makes no contribution to the integral. The contribution of the other may be evaluated by taking the loops round P and Q to be equal circles of small radius ϵ. The sum of the integrals round these circles is

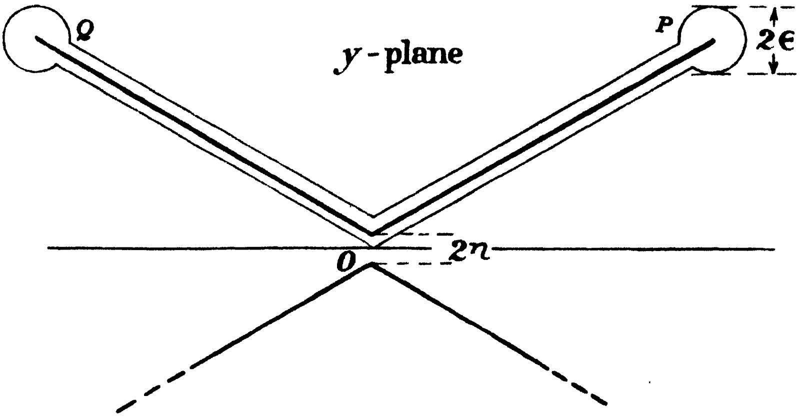


Figure 3. Cuts and contours for the integration of (20).

$$\frac{\pi\sin\theta}{(\cos\theta)^{\frac{1}{2}}}\left[\log\frac{\sin\theta}{(\cos\theta)^{\frac{1}{2}}}-\log\epsilon\right]. \qquad \ldots\ldots(21)$$

On the straight contours OP, PO write $y=\rho(\cos\theta)^{-\frac{1}{2}}e^{\frac{1}{2}i\theta}$, and the limits of ρ are 0 and $1-\epsilon(\cos\theta)^{\frac{1}{2}}$.

The integrand on OP exceeds the integrand at the neighbouring point on PO by

$$\xi(\rho)=\pi\frac{2[\cot\theta+i\rho^2/(\rho^2-1)]\cos\frac{1}{2}\theta+[(\rho^2-1)\cot^2\theta+2i\rho^2\cot\theta-(\rho^4+1)/(\rho^2-1)]\sin\frac{1}{2}\theta}{(\rho^2-1)\cot^2\theta+2i\rho^2\cot\theta-(\rho^2+1)}$$

and the sum of the integrals along these two contours OP, PO is

$$\frac{e^{\frac{1}{2}i\theta}}{(\cos\theta)^{\frac{1}{2}}}\int_0^{1-\epsilon(\cos\theta)^{\frac{1}{2}}}\xi(\rho)d\rho.$$

The sum of the integrals along OQ and QO is the complex conjugate of this. In the limit of small ϵ the sum of these four integrals is $\pi\sin\theta(\cos\theta)^{-\frac{1}{2}}$ times

$$1+\log\tan\tfrac{1}{2}\theta+\log\epsilon+\tfrac{1}{2}\log\cos\theta-\log 2. \qquad \ldots\ldots(22)$$

On adding to this the contribution of the small circular contours, the terms in $\log\epsilon$ cancel, and the integral in (20) becomes $\pi\sin\theta\,(\cos\theta)^{-\frac{1}{2}}F$, where

$$F=1+\log\tan\tfrac{1}{2}\theta+\log\sin\theta-\log 2. \qquad \ldots\ldots(23)$$

The contribution to the total energy of the forces acting across the plane P is now

$$-\mu d^2F/4\pi(1-\sigma). \qquad \ldots\ldots(24)$$

(b) *Potential energy of attraction between A and B*

The potential energy, referred to the homogeneously strained crystal, is

$$\frac{\mu d^2}{4\pi^2}\int_{-\infty}^{\infty}[\cos\tfrac{1}{2}\theta-\cos 2\pi\phi/d]\,dx. \qquad \ldots\ldots(25)$$

On substituting for ϕ and again changing the variable, as in equations (19) and (20), this becomes

$$\frac{\mu d^2(\cos\theta)^{\frac{1}{2}}}{4\pi^2(1-\sigma)\sin\theta}\int_{-\infty}^{\infty}[\cos\tfrac{1}{2}\theta-\cos 2\chi]\,dy. \qquad \ldots\ldots(26)$$

The integrand has simple poles in the upper half-plane at $y=\pm(\cos\theta)^{-\frac{1}{2}}e^{\pm\frac{1}{2}i\theta}$. On evaluating the residues the integral is easily shown to be $2\pi\sin\theta/(\cos\theta)^{\frac{1}{2}}$, and the contribution of the attraction between A and B to the energy is

$$\mu d^2/2\pi(1-\sigma). \qquad \ldots\ldots(27)$$

As was to be expected, this contribution remains finite when θ tends to zero.

(c) *Work done by external forces*

If a uniform shear stress $(\mu/2\pi)\sin\frac{1}{2}\theta$ is maintained on the external surface of the crystal during the formation of the pair of dislocations, the external forces do work

$$\frac{\mu}{2\pi}\sin\tfrac{1}{2}\theta\int_{-\infty}^{\infty}\left[\phi(x)-\frac{d\theta}{4\pi}\right]dx. \qquad \ldots\ldots(28)$$

As before, this may be written

$$\frac{\mu d^2\sin\frac{1}{2}\theta(\cos\theta)^{\frac{1}{2}}}{2\pi^2(1-\sigma)\sin\theta}\int_{-\infty}^{\infty}\cot^{-1}[(y^2-1)\cot\theta]\,dy. \qquad \ldots\ldots(29)$$

Integrating by parts, the integral in (29) becomes

$$2\tan\theta\int_{-\infty}^{\infty}y^2\,dy/[(y^2-1)^2+\tan^2\theta],$$

which may be evaluated from its residues as $2\pi\cos\frac{1}{2}\theta/(\cos\theta)^{\frac{1}{2}}$. The work done by the external forces, which is to be subtracted from the internal energy of the crystal in order to obtain the activation energy for the formation of a pair of dislocations, is, therefore,

$$\mu d^2/2\pi(1-\sigma). \qquad \ldots\ldots(30)$$

This is finite for small θ, because it represents the work done by forces of order θ acting through distances of order $1/\theta$.

The total activation energy is (24) + (27) − (30). For small θ this is approximately $\mu d^2 \log(2/\theta)/2\pi(1-\sigma)$. Its dependence on the distance between the dislocations agrees with that obtained by Koehler (1941).

To estimate the order of magnitude of this quantity, take $\mu = 4{\cdot}4\times10^{11}$ dynes/cm², $\sigma = \frac{1}{4}$, and for d take the distance of closest approach of two copper atoms, $2{\cdot}5\times10^{-8}$ cm. Then the coefficient $\mu d^2/2\pi(1-\sigma)$ is 6×10^{-5} erg/cm., or $1{\cdot}5\times10^{-12}$ erg per atom-pair. This is 1 electron volt for each atomic plane even for large strains of the order $\theta = 1$. For the practical elastic limits of metallic single crystals $\theta \simeq 10^{-3}$ and the energy is 7 electron volts per atomic plane.

§6. THE SHEAR STRESS REQUIRED TO MOVE A SINGLE DISLOCATION

In the previous section, the energy of a pair of dislocations has been calculated by treating the two halves of the crystal as elastic continua, and integrating over the interfaces A and B. To this approximation the energy of a single dislocation in a crystal free from external stress is independent of the position of the dislocation in the crystal. The dislocation is in neutral equilibrium, and will move under the smallest external shearing stress, since its motion causes a relative displacement of the half-crystals a and b in the x-direction. On the other hand if the atomic structure is taken into account, the energy of a dislocation must, in the absence of a stress, depend on its exact position, i.e. on whether the plane of symmetry passes through a row of atoms or not. Hence the dislocation will have a number of positions of stable equilibrium, and these will persist even under a stress until this exceeds a certain magnitude.

The stress required to move a dislocation may be estimated by assuming that each half-crystal a and b retains its form as the dislocation moves, so that the contribution analogous to (a) in §5 is independent of the position of the dislocation. The contribution analogous to (b) is to be evaluated by replacing the integral (25) by the corresponding sum over all atoms in the planes A and B. (It is not sufficient to sum over the atoms of plane A alone, for this would yield an interaction energy which was a periodic function of the displacement of the dislocation with period d. The geometrical form of the dislocation, and therefore its energy, is in fact restored by a displacement of only $\frac{1}{2}d$.) Since the energy is an energy of interaction between the atoms of A and the atoms of B, the sum must be halved to obtain the energy. Taking αd as the displacement of the centre of the dislocation from the position shown in figure 1, the energy becomes

$$\frac{1}{2}\frac{\mu d^2}{4\pi^2}\sum_{n=-\infty}^{\infty}[\text{const.}+\cos 2\{\tan^{-1}2(1-\sigma)(\alpha+\tfrac{1}{2}n)\}]. \qquad \ldots\ldots(31)$$

This may be transformed by using the relation

$$\sum_{-\infty}^{\infty} f(n) = \sum_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)\cos 2\pi xs\,dx \qquad \ldots\ldots(32)$$

into

$$\frac{\mu d^2}{8\pi^2}\int_{-\infty}^{\infty}[\text{const.}+\cos 2\{\tan^{-1}2(1-\sigma)(\alpha+\tfrac{1}{2}x)\}]\,dx$$

$$+\frac{\mu d^2}{4\pi^2}\sum_{s=1}^{\infty}\int_{-\infty}^{\infty}[\text{const.}+\cos 2\{\tan^{-1}2(1-\sigma)(\alpha+\tfrac{1}{2}x)\}]\cos 2\pi xs\,dx.$$

Writing

$$z=2(1-\sigma)(\alpha+\tfrac{1}{2}x)$$

reduces this to

$$\frac{\mu d^2}{8\pi^2(1-\sigma)}\int_{-\infty}^{\infty}[\text{const.}+\cos\{2\tan^{-1}z\}]\,dz$$

$$+\frac{\mu d^2}{4\pi^2(1-\sigma)}\sum_{s=1}^{\infty}\int_{-\infty}^{\infty}[\text{const.}+\cos\{2\tan^{-1}z\}]\cos 2\pi s\left(\frac{z}{1-\sigma}-2\alpha\right)dz.$$

The first term is an infinite constant independent of α, and the second reduces to

$$\frac{\mu d^2}{2\pi^2(1-\sigma)}\sum_{s=1}^{\infty}\cos 4\pi s\alpha\int_{-\infty}^{\infty}\cos\left(\frac{2\pi s z}{1-\sigma}\right)\frac{dz}{1+z^2}. \qquad \ldots\ldots(33)$$

The integral in (33) is equal to $\int_{-\infty}^{\infty} e^{ikz}dz/(1+z^2)$, which is readily shown by contour integration to be πe^{-k}, where $k=2\pi s/(1-\sigma)$. The energy is then

$$v=\frac{\mu d^2}{2\pi(1-\sigma)}\sum_{s=1}^{\infty} e^{-\frac{2\pi s}{1-\sigma}}\cos 4\pi s\alpha. \qquad \ldots\ldots(34)$$

The work done in an infinitesimal displacement δ of the dislocation is

$$\frac{\delta}{d}\frac{dv}{d\alpha}=-\frac{2\mu d\delta}{1-\sigma}\sum_{s=1}^{\infty} s e^{-\frac{2\pi s}{1-\sigma}}\sin 4\pi s\alpha.$$

The term in $s=1$ dominates this sum, and has a maximum value

$$\frac{2\mu d\delta}{1-\sigma}e^{-\frac{2\pi}{1-\sigma}}. \qquad \ldots\ldots(35)$$

The work done by an external shearing stress T in this displacement is $Td\delta$, and by comparison with (35) it follows that the smallest external stress which will cause a single dislocation to move continuously through the lattice is given by

$$T=\frac{2\mu}{1-\sigma}e^{-\frac{2\pi}{1-\sigma}}, \qquad \ldots\ldots(36)$$

whereas the shearing stress at which one lattice plane moves rigidly over another is $\mu/2\pi$. The ratio of the stress required to move a dislocation to the theoretical shear strength of a perfect lattice is

$$\frac{4\pi}{1-\sigma}e^{-\frac{2\pi}{1-\sigma}}. \qquad \ldots\ldots(37)$$

Since the width of the dislocation given by (12) is only half that given by (3) in the original note, the energy of a dislocation is more sensitive to its position in the lattice, and the ratio (37) is of a larger order of magnitude than the original estimate. Values are:

Poisson's ratio	0·2	0·3	0·4
Ratio of stresses	6×10^{-3}	2×10^{-3}	6×10^{-4}

Expression (37) differs from Peierls's expression not only in replacing $(1-\sigma)$ by $2(1-\sigma)$, but also in the form of the coefficient. This is because Peierls replaces the integrals (19) and (25) by the corresponding sums, whereas in (37) only (25) has been replaced by a sum. The order of magnitude of the result is the same on each approximation, the exponential factors agreeing. Any calculation of the effect of atomic structure which is based on the result (11) of the theory of a continuous medium is really inconsistent, and the answer is therefore undefined.

The shear stress T given by (36) would cause a uniform shear of the lattice through a small angle $\theta = T/\mu$, and this stress would just equal the mutual attraction of a pair of dislocations of unlike sign in the same glide plane at a distance apart of $2d/(1-\sigma)\theta$. To this approximation, a pair of dislocations in an unstressed crystal will coalesce if they are close together, but their attraction will not overcome the forces anchoring them to their places in the lattice if their separation exceeds

$$d e^{\frac{2\pi}{1-\sigma}}. \qquad \ldots\ldots(38)$$

Numerically this implies:—

Poisson's ratio	0·2	0·3	0·4
Critical separation	$2500d$	$8000d$	$35000d$

§7. DISCUSSION

To the approximation considered here, dislocations in a lattice have the following properties. The width of a dislocation is small, displacements comparable with the interatomic distance being confined to a few atoms. The energy required to form a pair of dislocations in a crystal under moderate shear stress is large, of the order of 10 electron volts for each atomic plane. The energy required to form even a short dislocation far exceeds the thermal energy of an atom at room temperature. A single dislocation will not move freely through a lattice under very small stresses, but will do so under shear stresses of the order of a thousandth of the theoretical shear strength of the perfect lattice. This figure is of the same order of magnitude as that commonly accepted as the elastic limit of a real single crystal. Two dislocations of opposite sign in the same glide plane attract each other with a force which at large distances is inversely proportional to the distance between them, and they will run together and coalesce unless they are separated by a distance of the order of 10000 lattice spacings. This spacing is of the same order as the size of the mosaic blocks of which real crystals are usually composed, and may represent the reason for the appearance of this characteristic length in the growth of real crystals.

It is, however, necessary to consider how far this approximate treatment of a simplified model can be expected to represent the properties of a real crystal. The first approximation is the use of a simple cubic lattice. Glide processes are best understood in face-centred cubic and hexagonal metals. In both cases glide takes place in a close-packed direction over close-packed planes, but the actual geometrical form of the dislocation in such lattices is not known or easy to visualize. It seems unlikely that this approximation should greatly affect the width or the energy of a dislocation, but the stress required to move a dislocation, and the critical separation of two dislocations, which depend very sensitively on the width of a dislocation, might be entirely different.

The next approximation is the assumption that the crystal is isotropic. It is easy to show (cf. Appendix) that the elastic modulus for shearing a cubic crystal in the direction $(1\bar{1}0)$ across a (111) plane is $4[s_{44}+\frac{1}{3}(s_{11}-s_{12}-2s_{44})]$. The anisotropy term $\frac{1}{3}(s_{11}-s_{12}-2s_{44})$ is only 20% of s_{44} for Cu, Ag and Au, so it seems likely that anisotropy will not be of great importance. Here again the stress required to move a dislocation depends critically on its width.

Probably the most serious assumption is the sinusoidal law of force (1). In the majority of metals, the ionic shells are in contact, and for many purposes it is satisfactory to regard their lattices as composed of hard spheres in contact. Dr. Orowan has pointed out in discussion that such an extreme model does not have an elastic modulus in the usual sense, because a finite force is required to produce an infinitesimal displacement of two neighbouring layers of atoms. If this crude picture is refined somewhat by considering the spheres to be held apart by forces which vary very much more rapidly with distance than the attractive forces, the work required to slide one layer of atoms a distance $\frac{1}{2}d$ over the next layer is still much less than μd^2, which is the order of magnitude corresponding to (1). It follows from the discussion of § 1 that if in fact the surface energy is much less than that corresponding to a sine law, the width of a dislocation will be greater than that calculated, since an increase in the width reduces its elastic energy without a corresponding increase in its surface energy. Its total energy is less (but could hardly be so much reduced that pairs of dislocations could be formed by thermal agitation) and it moves more easily.

The following argument, due to Professor Mott, shows that this correction is not large. Consider a two-dimensional model, which in its equilibrium state consists of close-packed cylinders (figure 4*a*). In the dislocated state the cylinders are arranged as in figure 4*b*.

Let the distance between the layers of atoms on either side of the slip plane be z, while each atom in the upper row is displaced horizontally a distance ϕ from its neighbour in the lower row. It is assumed in the argument that the only forces acting are a long-range force, the potential of which depends only on the volume of the crystal, and short-range repulsions between neighbouring ions. This is not correct, for the alkali metals, in which the short-range repulsion is small, have rigidities comparable with their bulk moduli. These neglected forces are such that the configuration of figure 4*a* is one of stable equilibrium; when the magnitude of the other forces is estimated from the observed rigidity of the crystal under small displacements it is therefore over-estimated. The neglected forces depend on long-range interactions alone, and are consequently smoothly varying functions of ϕ which cannot depart far from a sinusoidal form. The estimate of the departure of the law of force from a sinusoidal form which is made by neglecting these forces must exceed the true departure.

It will be convenient to represent the energy of the cohesive forces by an expression of the form Bz^2, that is to say, proportional to V^2, where V is the atomic volume. In fact the electrostatic energy is proportional to $V^{-\frac{1}{3}}$, and is opposed by the Fermi energy, which is proportional to $V^{-\frac{2}{3}}$. The constant B is determined by the condition that the long-range force should equal the short-range repulsion at the equilibrium interatomic distance. It may be shown that for larger values of V the deviation of the true energy from that given by the approximate expression Bz^2 is greatest when the Fermi energy is neglected, and that if the Fermi energy is neglected the approximate formula over-estimates the difference of the energies of figures 4 *a* and 4 *b* by less than 20%. The repulsive potential, which falls off rapidly with distance, is taken to be of the form $Ae^{-\beta r^2}$, where $\beta d^2 \gg 1$.

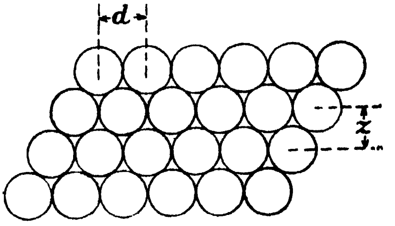


Figure 4 *a*. A normal crystal ($\phi = \frac{1}{2}d$).

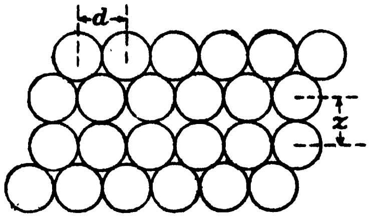


Figure 4 *b*. A dislocated crystal ($\phi = 0$).

The energy per atom is then

$$W = Bz^2 + A \sum_{n=-\infty}^{\infty} e^{-\beta[(nd+\phi)^2+z^2]}. \qquad \ldots\ldots(39)$$

For each value of ϕ, the equilibrium value of z is given by $\partial W/\partial z = 0$, which leads to

$$W = \frac{B}{\beta}\left[1 + \log\left\{\frac{\beta A}{B}\sum_n e^{-\beta(nd+\phi)^2}\right\}\right]. \qquad \ldots\ldots(40)$$

For figure 4 *a*, $\phi = \frac{1}{2}d$, and, neglecting the repulsions between all except nearest neighbours,

$$W_a = \frac{B}{\beta}\left[1 + \log\frac{2\beta A}{B} - \tfrac{1}{4}\beta d^2\right].$$

Similarly

$$W_b = \frac{B}{\beta}\left[1 + \log\frac{\beta A}{B}\right].$$

The energy difference is

$$W_b - W_a = \frac{B}{4\beta}[\beta d^2 - 4\log 2]. \qquad \ldots\ldots(41)$$

Near the position of stable equilibrium (40) may be written approximately

$$W = \frac{B}{\beta}\left[1 + \log\frac{\beta A}{B} + \log\left\{e^{-\beta(d-\phi)^2} + e^{-\beta\phi^2}\right\}\right].$$

This gives

$$\frac{dW}{d\phi}=2B\frac{(d-\phi)e^{-\beta(d-\phi)^2}-\phi e^{-\beta\phi^2}}{e^{-\beta(d-\phi)^2}+e^{-\beta\phi^2}},$$

which vanishes for $\phi=\frac{1}{2}d$, and when $\phi=\frac{1}{2}d$,

$$\frac{d^2W}{d\phi^2}=B(\beta d^2-2). \qquad \ldots\ldots(42)$$

For a sinusoidal law of force of the form

$$W=P+Q\cos 2\pi\phi/d$$

the relations corresponding to (41) and (42) are

$$W_b-W_a=2Q \qquad \ldots\ldots(43)$$

$$\frac{d^2W}{d\phi^2}=\frac{4\pi^2 Q}{d^2}. \qquad \ldots\ldots(44)$$

If now Q is determined by a comparison of (42) and (44), which is the method used in §2, its value is found to be

$$Q=Bd^2(\beta d^2-2)/4\pi^2,$$

and the energy of each atom in the dislocated plane exceeds that of a normal atom by the amount given by (43), namely,

$$W_b-W_a=Bd^2(\beta d^2-2)/2\pi^2. \qquad \ldots\ldots(45)$$

This value (45) exceeds the true value (41) by a factor

$$\frac{2\beta d^2}{\pi^2}\cdot\frac{\beta d^2-2}{\beta d^2-4\log 2}. \qquad \ldots\ldots(46)$$

For very hard ions $\beta d^2 \gg 1$ and the factor (46) is large. The value of (46) for copper may be estimated from the values of dW/dr and d^2W/dr^2 given by Fuchs (1935). Here W is the repulsive energy, which in this calculation is assumed to be given by $W=Ae^{-\beta r^2}$. It follows that

$$\beta r^2=\tfrac{1}{2}-r\frac{d^2W}{dr^2}\Big/2\frac{dW}{dr},$$

and Fuch's numerical values lead to $\beta r^2=20{\cdot}3$ at the equilibrium distance $r=d$. The value of the factor (46) is then 4·25. The sinusoidal law (1) therefore overestimates the energy of a surface of misfit by a factor of less than 4·25. The corresponding errors in estimating the size and energy of a dislocation are at most factors of about 2.

It seems unlikely that the purely mathematical approximations should greatly influence the result. The most dangerous assumption seems to be that of §6, which is that the displacements of the individual atoms are given exactly by (12) for all positions of the dislocation in the lattice. A more detailed investigation might show that the second approximations for the displacements of the atoms in the extreme positions of the dislocation in the lattice (corresponding to $\alpha=0$ and $\alpha=\frac{1}{4}$ in (31)) lead to differences in the energies of these two positions which cannot be neglected in comparison with the very small difference calculated by using (12) in each case. It is difficult to see how to obtain such a second approximation, or to decide on general grounds whether the reduction in energy corresponding to the second approximation would be greater for the stable position ($\alpha = 0$) or the unstable position ($\alpha = \frac{1}{2}$).

ACKNOWLEDGMENTS

The writer's thanks are due to Professor Peierls for an explanation of the methods used in deriving the results previously published, and for permission to quote sections of his paper. He is also indebted to Professor Mott for valuable discussions, and to Dr. E. H. Linfoot for pointing out some of the questions of convergence involved in the derivation of equation (10).

APPENDIX

THE CONTRIBUTION OF ANISOTROPY TO THE SHEAR MODULUS IN THE SLIP DIRECTION

Consider the stress system which, referred to the principal axes of a cubic crystal, is represented by

$$\begin{bmatrix} 2 & 0 & 1 \\ 0 & -2 & -1 \\ 1 & -1 & 0 \end{bmatrix} \times \frac{S}{6^{\frac{1}{2}}}. \qquad \text{......(A 1)}$$

This is easily shown to represent a shearing stress S across the plane (111) in the direction $(1\bar{1}0)$.

The corresponding strain in a material of cubic symmetry is

$$\begin{bmatrix} 2(s_{11}-s_{12}) & 0 & 2s_{44} \\ 0 & -2(s_{11}-s_{12}) & -2s_{44} \\ 2s_{44} & -2s_{44} & 0 \end{bmatrix} \times \frac{S}{6^{\frac{1}{2}}}. \qquad \text{......(A 2)}$$

The component of shear strain in the direction corresponding to the stress is

$$2Ss_{44} + \tfrac{2}{3}S(s_{11}-s_{12}-2s_{44}),$$

and the angle of shear is twice this.

REFERENCES

Burgers, J. M., 1940. *Proc. Phys. Soc.*, **52**, 23.
Dehlinger, V., 1929. *Ann. Phys., Lpz.*, **2**, 749.
Fuchs, K., 1935. *Proc. Roy. Soc.*, A, **153**, 622.
Koehler, J. S., 1941. *Phys. Rev.*, **60**, 397 (equation 18).
Love, A. E. H., 1944. *Elasticity* (Cambridge : The University Press) (§ 144).
Peierls, R., 1940. *Proc. Phys. Soc.*, **52**, 34.
Taylor, G. I., 1934. *Proc. Roy. Soc.*, A, **145**, 362.
Titchmarsh, E.C., 1937. *Fourier Integrals* (Oxford : The Clarendon Press), Chap. V.

A NOTE ON THE EFFECT AT THE CATHODE OF AN ARC BETWEEN COPPER ELECTRODES

By MAURICE MILBOURN,
Imperial Chemical Industries Ltd., Metals Division

MS. received 19 *September* 1946

ABSTRACT. **Observations on burning arcs and on arced electrodes have indicated that melting of a copper cathode does not necessarily take place, and that selective distillation of impurities occurs when melting is induced by the presence of a metal having powerful reducing properties. Volatilization of copper from the cathode appears to be effected through the formation of cuprous oxide.**

§1. INTRODUCTION

In a normal copper arc, as used for spectrographic analysis, volatilization takes place at a greater rate from the negative than from the positive electrode. Selective distillation of impurities from the cathode does not occur to any great extent, but elements with low boiling points may be removed preferentially if free oxidation of the negative electrode is hindered by the presence of a metal having powerful reducing properties. At the same time, the rate of volatilization of copper is reduced (Milbourn, 1943).

It has been suggested (Milbourn, 1944) that impurities distil out of the electrode material only when an appreciable volume of metal is molten, and that the absence of selective distillation indicates that no such melting is taking place. This point of view cannot easily be reconciled, however, with generally accepted theories of the arc discharge, which postulate that the cathode is at or near the boiling point of the metal. For instance, it has been estimated by von Engel and Steenbeck (1934) that a copper cathode attains a temperature of about 2000° K. At the same time, layers of oxide seem to be necessary for the maintenance of a normal arc, so that a reasonable energy balance may be established (von Engel, 1935).

§2. EXPERIMENTAL OBSERVATIONS

In order to obtain information on the volatilization of material, the degree of melting, and the distribution of oxide layers in copper cathodes, observations have been made on arcs while they were burning, and on electrodes after they had been subjected to an arc 2 to 3 mm. in length and carrying a current of 5 amp. for 30 to 60 seconds, conditions which could very well be used for analytical purposes.

Examination of a burning copper arc with a low-power binocular microscope, using a suitable filter, showed that the cathode spot is more or less surrounded by a rim of molten material, but that it leaves behind it, as it travels, an area which is free from such material. It appears, in fact, that the cathode spot travels towards regions where molten material has collected and vaporizes it.

Sections cut through the small area on the negative electrode where the arc was striking when it was extinguished have been examined microscopically, some of the results being illustrated in the accompanying photographs. Figure 1 shows a section of a ¼ in. diameter pure copper rod, on which the arc passed to a pyramidal point. The shallow depression in the surface of the specimen is a section of the area where the cathode spot was acting when the arc was extinguished, and although there is no apparent layer superimposed on the copper in the depression, there is a rim of bluish-coloured cuprous oxide on each side of it. These rims form the molten material observed on the surface of the electrode while the arc was running. The fact that the crystal structure of the copper is undisturbed up to the point of striking of the arc indicates that very little, if any, melting of the metal itself has taken place, and that its temperature was probably considerably below 1000° c. Figure 2 is taken from a similar section through an arced sample of copper strip, about 0·020 in. thick. The oxide rim can again be seen, but in this case the smaller dimensions of the sample have allowed the material to become hotter, although melting has not occurred to any detectable extent.

A strip of 93/7 copper-aluminium alloy, illustrated in figure 3, shows very considerable melting and little oxidation under similar conditions. It is clear that selective distillation would be expected from this sample to a much greater degree than from the copper samples illustrated in figures 1 and 2. The observations on melting thus confirm those made previously on the relative amounts of selective distillation from copper and from small samples containing a powerful reducing agent. The presence of aluminium has prevented oxidation of the copper, so that the heat of vaporization of copper oxide, as well as the heat of formation of aluminium oxide, are made available for melting the basis metal.

By moving a flat copper surface horizontally at about one inch per second, while an arc was striking it vertically, a continuous streak was obtained showing the instantaneous state of the cathode, illustrated in figure 4.

The central region of the arc streak has well-defined, although irregular, edges, and the area between them is covered with a very thin layer of cuprous oxide, which is readily soluble in dilute hydrochloric acid, together with a few strings of globules of the same material. Both of these are firmly adherent, but the material outside this central region can be easily removed with dry cotton wool. The evidence is that the cuprous oxide layer is formed while the arc is passing, and that it collects firstly into small globules and then into comparatively thick layers at the edge of the arc, from whence it is volatilized. It may be noted that cupric oxide decomposes into cuprous oxide at 1200° c., and that the latter melts at 1235° c., eventually decomposing into copper and oxygen at 1800° c. The equilibrium of various compounds at high temperature may therefore play a very important part in the functioning of an arc.

§ 3. CONCLUSIONS

These observations show that melting of the negative electrode does not necessarily take place during the running of a normal arc between copper electrodes and, if it does, it can be an extremely localized effect. Selective distillation is liable to occur when the basis metal is molten, but volatilization of copper is

Figure 1. Section through cathode spot on copper rod. × 100.

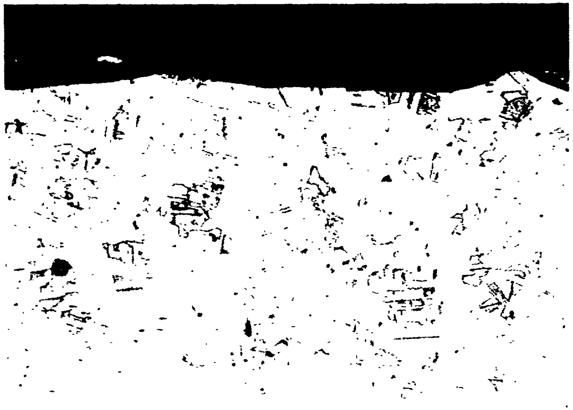


Figure 2. Section through cathode spot on copper strip. × 100.

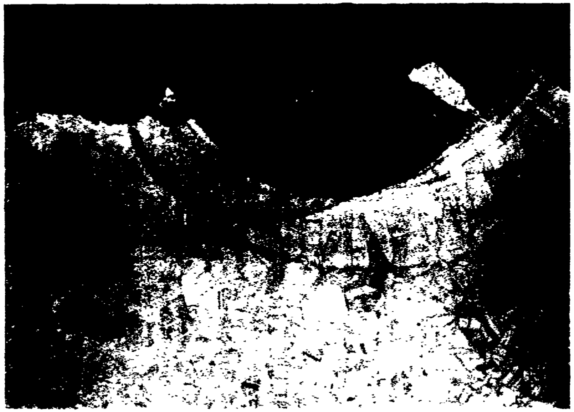


Figure 3. Section through cathode spot on aluminium bronze. × 20.

Figure 4. Arc streak on copper cathode. × 65.

effected through the formation of cuprous oxide. These observations have a direct bearing on the spectrographic analysis of copper and its alloys by an arc technique, but they may also be of value in connection with other types of spectrographic source, which include arc-like phases.

§4 ACKNOWLEDGMENT

The author's thanks are due to Dr. Maurice Cook for his helpful advice and interest in the work.

REFERENCES

von Engel, A. and Steenbeck, M., 1934. *Elektrische Gasentladungen*, vol. ii, p. 136.
von Engel, A., 1935. *Wiss. Veröff. SiemensWerken*, **14**, 38.
Milbourn, M., 1943. *J. Inst. Metals*, **69**, 441–463.
Milbourn, M., 1944. *J. Inst. Metals*, **70**, 299.

DAMPING CAPACITY, STRAIN HARDENING AND FATIGUE

By R. F. HANSTOCK,
High Duty Alloys Ltd.

MS. received 17 *August* 1946

ABSTRACT. An electromagnetic method of exciting torsional resonance vibrations is described.

For some alloys of aluminium, notably binary alloys containing 5% and 11% of magnesium, vibrational strains of sufficient magnitude to cause fatigue cracks can be developed by this method at frequencies of the order of one kilocycle per second. During vibration, measurements are made of the damping capacity of the specimens to provide information concerning strain hardening and energy conversion within the specimens up to the time when failure occurs by fatigue.

Fatigue failure of the two binary alloys containing magnesium is shown to be preceded by strain hardening. The endurance of individual specimens is related to the amount of irreversible strain imposed per cycle, failure occurring when the cumulative internal strain approaches a limiting value.

§1. INTRODUCTION

When a polycrystalline metal is subjected to alternating stresses, the material may fail by fracture after a certain number of stress cycles even though the maximum stress involved is considerably lower than the static stress required to cause rupture. Described in general terms as fatigue, this phenomenon is investigated conventionally by subjecting a specimen of the metal to an alternating stress (frequency normally about 100 cycles per second), the number of stress cycles required to cause fracture being recorded. A series of tests of this type, over a range of maximum values of the alternating stress, gives a fatigue curve for the material.

Strain hardening is a result of plastic or irreversible straining of the metal and is characterized by an increase in the stress necessary to produce a defined amount of plastic strain, i.e. as an increase in the limit of proportionality or proof stress. During strain hardening a part of the work done is stored in the metal and subsequently may be released by annealing.

Damping capacity, or internal friction, refers to the capacity of a solid to convert vibrational energy to some other form. It is defined normally as $\Delta E/E$, where ΔE is the energy absorbed and/or dissipated during each cycle of the vibration, and E is the total mechanical energy stored within the material for defined conditions of vibration.

Fatigue and strain hardening are the results of energy conversion, whilst damping capacity expresses the ability of the material to cause this conversion. The object of this paper is to show the relation of damping capacity to strain hardening and fatigue by referring to an experimental examination of two alloys of aluminium and magnesium.

This work became possible as a result of the development of a method of measuring damping capacity during vibration tests on specimens at strains within the fatigue range. The method is described in an earlier paper (Hanstock and Murray, 1946) which also records that damping capacity is a function of the maximum vibrational surface strain, ϕ, of the metallurgical state and of the vibrational history of the specimen. It is also known, qualitatively, that $\Delta E/E$ is sensitively dependent on the temperature of the specimen when a certain temperature, probably characteristic of the material, is exceeded. In the present investigation, the temperature of the specimens was kept within the region where $\Delta E/E$ is insensitive to variations of temperature.

§2. FORM OF SPECIMEN AND METHOD OF EXCITATION

The specimens have the form of the solid of revolution, about the vibration axis, XY, of the section shown in figure 1. The mode of vibration about XY is torsional, PQ being the nodal plane. Sections AB and CD are "inertias" and section BC is the elastic member of the system. The natural frequency of such a specimen (of alloys consisting mainly of aluminium) is about 960 cycles per second, and this is the frequency at which the fatigue tests are made.*

The specimen, S, figure 2, is suspended freely by a fine steel wire, W (0·018 cm. diameter and approximately 20 cm. long), the vibrational axis being vertical. At the junction of the wire and the specimen there is a small mirror, M, which reflects a beam of light from the optical system, L, on to a graduated scale (not shown). From the spread of the beam of light on the scale, the amplitude of vibration and the maximum surface strain on the elastic section of the specimen may be calculated. The mirror is brought into a suitable position for reflection by rotating the wheel, R. Vibrations at resonance frequency are excited by applying a simple harmonic couple electromagnetically to the upper end of the specimen. No mechanical coupling is involved, so that conversion of vibrational energy, other than as a result of the internal friction of the specimen, is very small.

---

* The actual frequency is measured during the test by means of a calibrated beat-frequency oscillator.

The following is a brief description of the method of excitation, but reference should be made to an earlier paper (*loc. cit.*) for a more detailed description of the electrical equipment. Two coils, C, figure 2, surround but do not touch the lower inertial section of the specimen. A magnetic field, originating from the permanent magnet, P, is so disposed in relation to these coils that, when the inertia is in alternating angular motion about the vibration axis, an induced current flows within this part of the specimen and causes an alternating e.m.f. to be developed in the coils. This e.m.f., which is of the same frequency as, and is proportional to the amplitude of the motion of, the specimen, forms the signal input of an amplifying system. The output of the amplifier (50 watts maximum) passes to coils, E, surrounding but not touching the upper inertial section of the specimen. The current induced in this inertia reacts with a magnetic field (produced by energizing the coil, F) to generate an alternating couple acting on the specimen.

Thus the whole system comprises an electromechanical oscillator, the frequency of which is determined by the dimensions and nature of the specimen, whilst the amplitude of the mechanical vibration depends on the damping capacity of the specimen and on the power supplied by the amplifier. The lower coil system, or detector, is magnetically shielded; part of the shielding has been removed in figure 2 to allow the coils to be seen. The upper coil system, or exciter, is air cooled through pipes, A, to prevent the temperature of the specimen rising to within the region where $\Delta E/E$ becomes dependent on temperature.

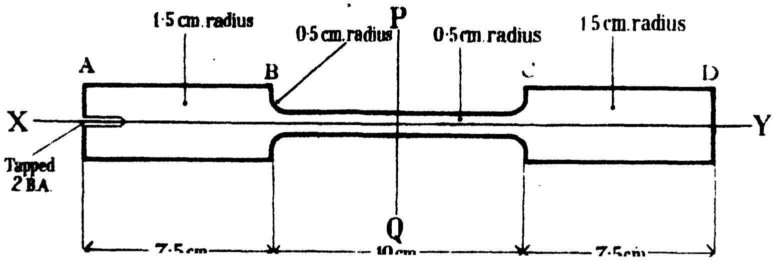


Figure 1. Section of specimen in plane containing vibration axis X, Y.

§3. MEASUREMENT OF DAMPING CAPACITY

For suitably constructed exciter and detector coils it may be shown that $\Delta E/E$ is directly proportional to I/V (to within 3% for aluminium alloys), where I is the R.M.S. current in the exciter and V is the R.M.S. voltage developed in the detector. For aluminium alloys it has been shown (*loc. cit.*) that $\Delta E/E$ is substantially independent of variation of the maximum surface shear strain for strains generally up to 1×10^{-4} and very often, depending on the composition and metallurgical state of the alloy, up to strains considerably in excess of this value. Consequently it is permissible to determine $\Delta E/E$ within this range of strain by measuring the logarithmic decrement, λ, of free decay of the vibration. Provided that $\Delta E/E$ is small, $\Delta E/E = 2\lambda$. One absolute determination of $\Delta E/E$ from a measurement of the half-amplitude decay time at low strains is sufficient to allow the proportionality

$$\Delta E/E \propto I/V$$

to be used for the determination of $\Delta E/E$ from I/V at higher surface strains during steady vibration of the specimen.

The half-amplitude decay time at low strains is determined by means of a stop-watch, the decay of amplitude being observed as a decay of the detector voltage. For specimens of the type under consideration, the half-amplitude decay time is about 29 seconds, which corresponds to a damping capacity of 5×10^{-5}. In order to simplify observations during fatigue tests, the ratio I/V is determined directly by potentiometric comparison of the detector voltage and of the voltage drop across a one-ohm resistance carrying the exciter current.

Having determined $\Delta E/E$, the following quantities may be calculated:—
Energy dissipated per cycle

$$\Delta E = \Delta E/E \,.\, \tfrac{1}{4} n v \phi^2 \text{ ergs.} \qquad \ldots\ldots(1)$$

where n = modulus of rigidity (dyne/cm.^{2});
v = volume of the elastic section of the specimen;
ϕ = peak value of the alternating surface strain on the elastic section of the specimen (radians).

Maximum value of the alternating couple on the specimen

$$G = \frac{\Delta E}{E} \,.\, \frac{n r^3 \phi}{2} \text{ dyne cm.} \qquad \ldots\ldots(2)$$

where r = radius of the elastic section of the specimen (cm.).

The values of $\Delta E/E$, as determined, represent the total damping capacity per cycle and include energy dissipations external to the specimen (e.g. losses through and in the specimen wire, frictional losses at the junction of the wire and specimens, air friction losses and losses due to electromagnetic damping).

The sum of these losses can be estimated experimentally, and in the present work is found to be represented by a damping capacity of 2×10^{-5}, which is independent of the amplitude of the vibration.

§4. THE ALLOYS INVESTIGATED

It will be appreciated that with an electrical amplifying system of limited output, the maximum vibrational strain which can be developed in the mechanical vibrator will depend on the damping capacity of the material. During a series of experiments with various alloys of aluminium, it was found that binary alloys containing 5% and 11% of magnesium had a low and practically constant damping capacity for surface shear strains up to about 15×10^{-4} and 30×10^{-4} respectively, above which values the damping capacity began to increase. This relation between damping capacity and strain is typical of aluminium alloys, but the range of strains over which the damping capacity remains small and constant is unusually extensive for the alloy containing 11% magnesium. Further, these alloys are known to be capable of appreciable strain-hardening and, as the results to be given later will show, can be strain-hardened by vibration, thus causing the damping capacity at any strain, greater than the respective values given above, to decrease during vibration. It appeared, therefore, that it would be possible to excite vibrations of sufficient amplitude to produce fatigue cracks in these alloys, even though the maximum power available did not exceed 50 watts.

Normally, the metallurgical state of such alloys would be made definite by raising them to a temperature of 430° c. for sufficient time to ensure complete solution of the Al_2Mg_3 compound and by rapid cooling to retain this compound in solid solution; alternatively they might be cooled slowly from 430° c. to allow the precipitation of Al_2Mg_3. The separation of Al_2Mg_3 causes a considerable reduction in ductility. In the present investigation the metallurgical state of the material was not of the first importance. Primarily it was necessary to obtain a number of specimens all in a similar stable condition. The specimens were taken from 3½ in. diameter extruded bar, stored for two years after extrusion so that no marked differences in the metallurgical states of the individual specimens would be expected. Metallographic examination of the alloys showed that the Al_2Mg_3 was in solution in the 5% alloy but not entirely so in the 11% alloy. A few specimens of the 11% alloy, heat-treated to obtain complete solution, were found to behave not very differently from the alloy as extruded, when subjected to the vibration tests.

The specimens were prepared by careful machining to the dimensions shown in figure 1, from pieces sawn off-centre from the 3½ in. diameter bar. After machining, the central section of each specimen was polished.

§ 5. METHOD OF EXAMINATION

In conventional fatigue tests, the specimen is subjected to stress or strain alternations between certain defined limits, the cyclic process being continued until a fatigue fracture occurs. It is not convenient to adopt these conditions for the operation of the electromechanical system owing to the limited output of the amplifier.

In many cases the damping capacity of the specimen changes during vibration. If it increases, the output of the amplifier must be increased to maintain a constant amplitude of vibration and eventually the maximum output may be insufficient to maintain the desired amplitude. If the amplitude decreases during vibration, as is the case for the alloys now being considered, the output must be decreased to maintain a constant amplitude, and consequently the maximum available power will not be employed except for short periods at the commencement of the test; under such conditions the equipment will not be used efficiently for production of fatigue.

In practice, the amplifier is operated to maintain a constant value of the R.M.S. current output whilst testing any one specimen. This operating condition is such that the peak value of the alternating couple applied to the specimen remains constant throughout the test.

The course of a typical test is as follows:—The specimen is vibrated at an amplitude corresponding to a maximum surface strain of $0{\cdot}75 \times 10^{-4}$ and the I/V ratio is measured. The strain is then increased to $1{\cdot}5 \times 10^{-4}$, excitation being stopped when this value is reached. As the amplitude decays, the time is measured for the strain to fall from $1{\cdot}0 \times 10^{-4}$ to $0{\cdot}5 \times 10^{-4}$. From this decay time, the damping capacity corresponding to an average strain of $0{\cdot}75 \times 10^{-4}$ is calculated. The specimen is then vibrated at the selected value of the exciter current and observations are made of the amplitude and of the I/V ratio as the test proceeds, until either the specimen fails by fatigue or the number of vibration cycles exceeds 10^8. The frequency of the specimen is determined periodically during the test.

§6. RESULTS

Alloy containing 11% *magnesium*

The experimental observations on the 11% alloy are summarized in figures 3 and 4, which show respectively, the change in surface shear strain (i.e. amplitude) and the change in damping capacity during vibration up to fracture, or to more than 10^8 cycles, of several specimens operated on by couples of peak values ranging from 4 gm./cm. to 100 gm./cm.

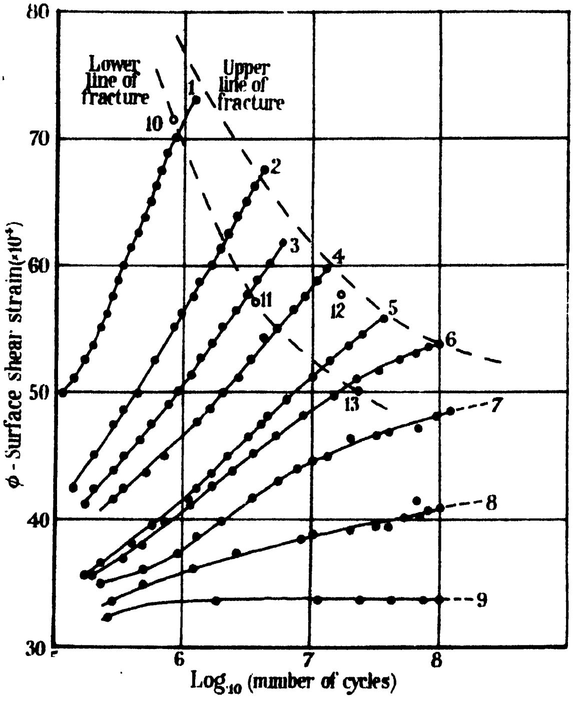


Figure 3. Variation of strain (amplitude) during vibration of an alloy of aluminium containing 11% Mg.

Increase in strain and decrease in damping capacity occur during vibration, the rate of change depending on the magnitude of the alternating couple. The formation of a fatigue crack is indicated sharply by a decrease in amplitude, an increase in damping capacity, and a fall in frequency of up to 100 cycles per second, depending on the size and disposition of the crack. Figure 5 shows a typical fracture exhibiting the usual characteristics of failure by fatigue. The specimens do not fracture completely during vibration, but can be broken easily afterwards, since the crack normally extends at least to the vibration axis. The rate at which the crack is formed is fairly high, a crack of the dimensions shown in figure 5 occurring in about 2 minutes (10^5 cycles).

The upper broken curve in figure 3 is a type of fatigue curve but differs from the conventional type in that the conditions it represents are not constant stress or

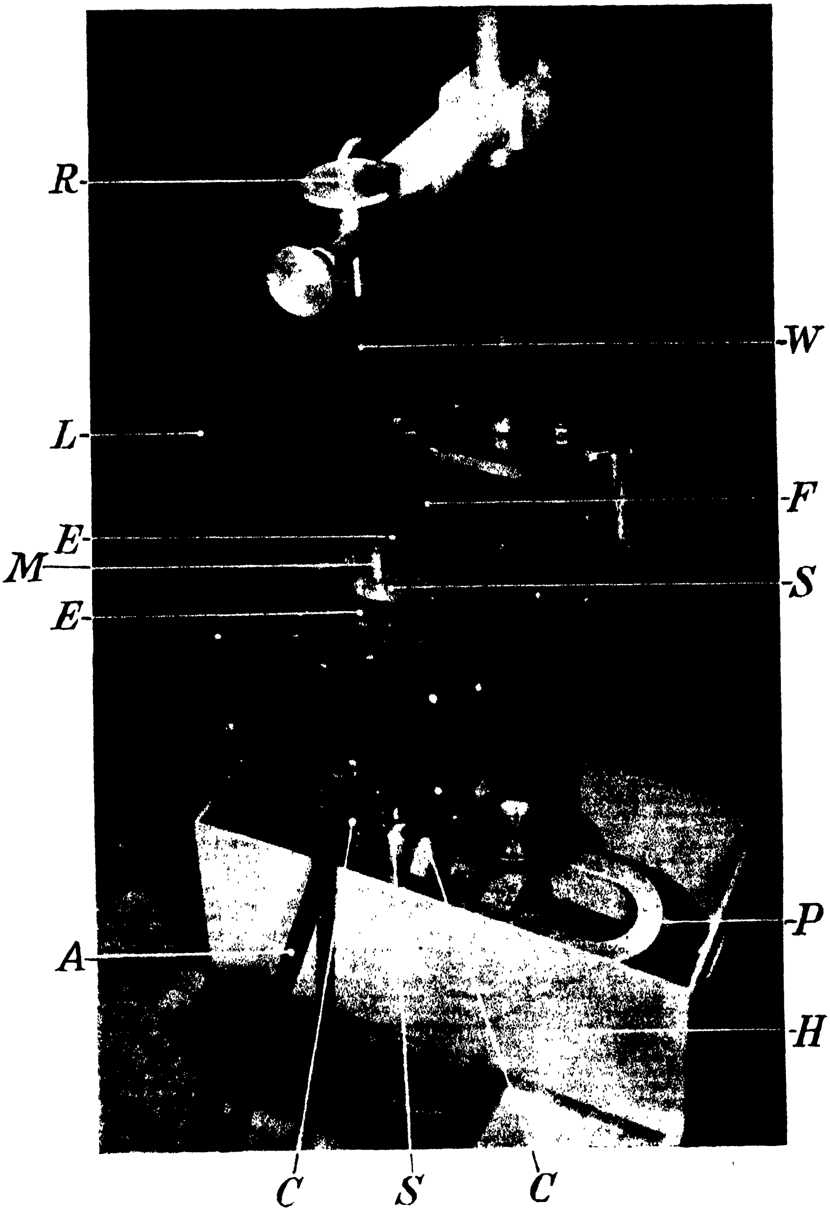


Figure 2. Exciter and detector units for torsional vibrations.

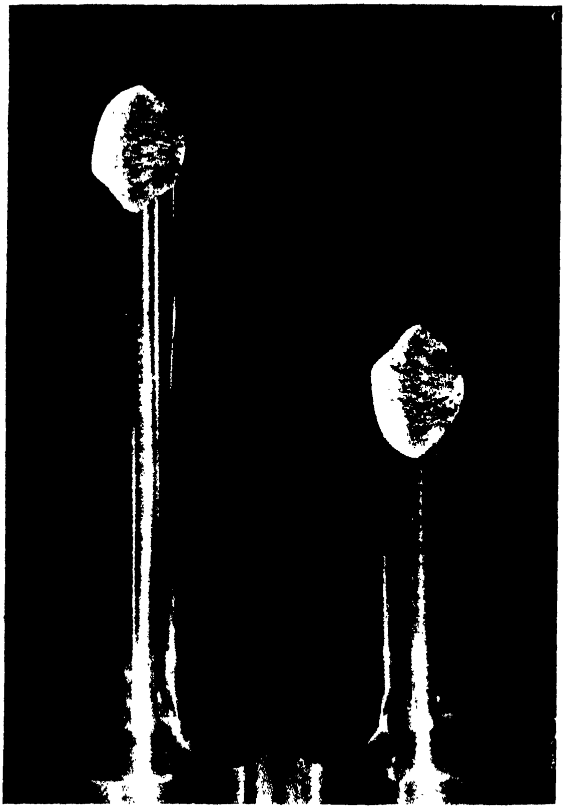


Figure 5. Fatigue fracture of aluminium alloy containing 11% Mg.

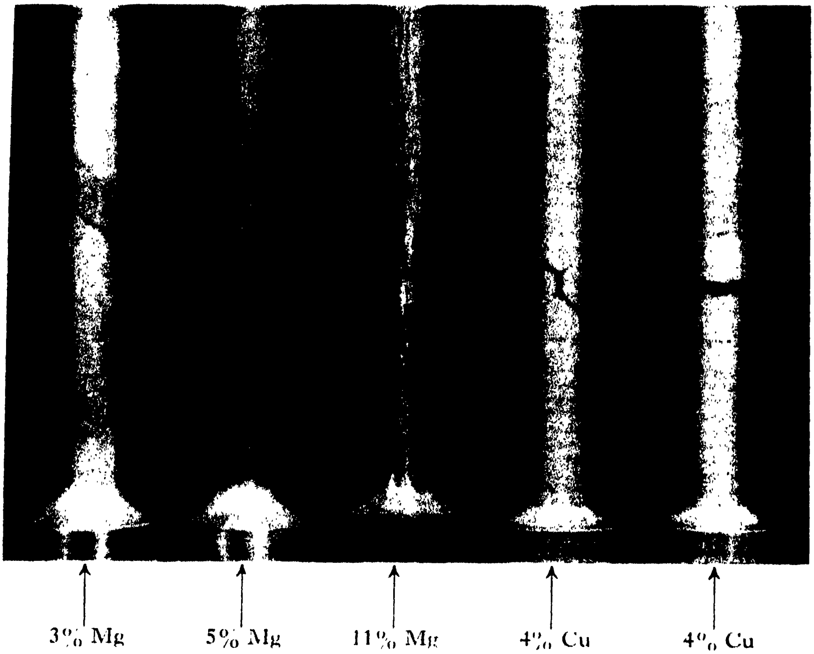


Figure 8. Fatigue cracks in aluminium alloys.

strain for each test but constant value of the alternating couple applied to a resonant system.

Further observed and derived data are given in table 1.

Table 1

Specimen number	Frequency (c.p.s.)	Couple (gm. cm.)	Total no. of cycles	$\Sigma_0^N \Delta E$ (ergs)	Mean ΔE (ergs)
1	930	100	$1{\cdot}23 \times 10^6$	$2{\cdot}29 \times 10^{10}$	18650
2	954	40	4·3	3·10	7200
3	965	25	6·0	2·30	3830
4	959	15·5	13·0	3·00	2310
5	973	11·2	37·1	5·15	1390
6	969	8·7	101	10·9	1080
7	975	7·4	124*	10·2	825
8	928	6·0	103*	5·74	556
9	962	4·0	103*	3·07	298

* Specimen not fractured.

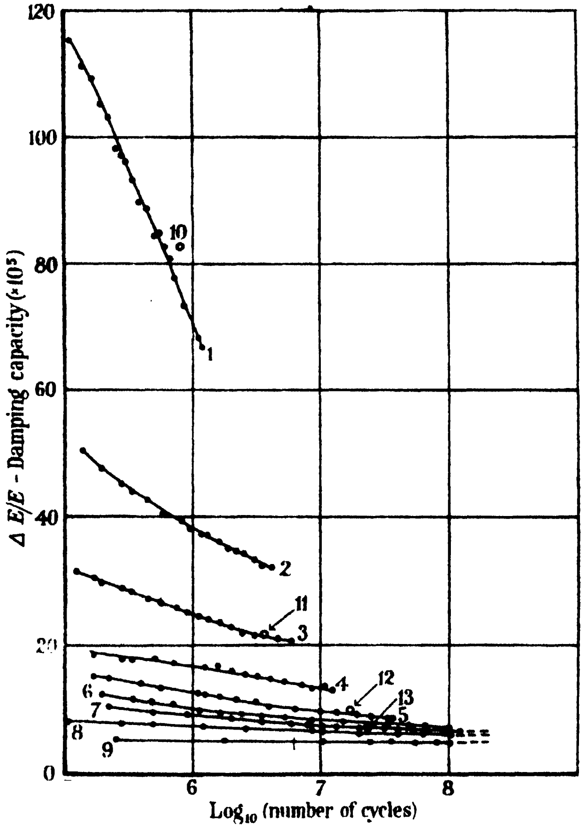


Figure 4. Variation of damping capacity during vibration of an alloy of aluminium containing 11% Mg.

From figures 3 and 4, graphs may be drawn showing the variation of ΔE with N, and from these graphs the total energy dissipated at the end of the test may be estimated. This energy is shown as $\Sigma_0^N \Delta E$ in table 1; the mean value of

ΔE (energy converted per cycle) given in this table is $1/N\,\Sigma_0^N \Delta E$. In addition to the specimens Nos. 1 to 9, for which the complete observations during test are given in figures 3 and 4, the end points of the test (corresponding to fatigue failure) are given in these figures for four more specimens, Nos. 10 to 13. Three of these specimens (Nos. 10, 11 and 13) failed prematurely but are included to indicate the lower limits of endurance observed in tests on this alloy.

The dependence of damping capacity on strain after a given number of cycles and for particular values of the alternating couple may be obtained from figures 3 and 4. For example, this dependence is shown in figure 6 at 0·25, 1·0, 2·5, 10, 25 and 100 million cycles.

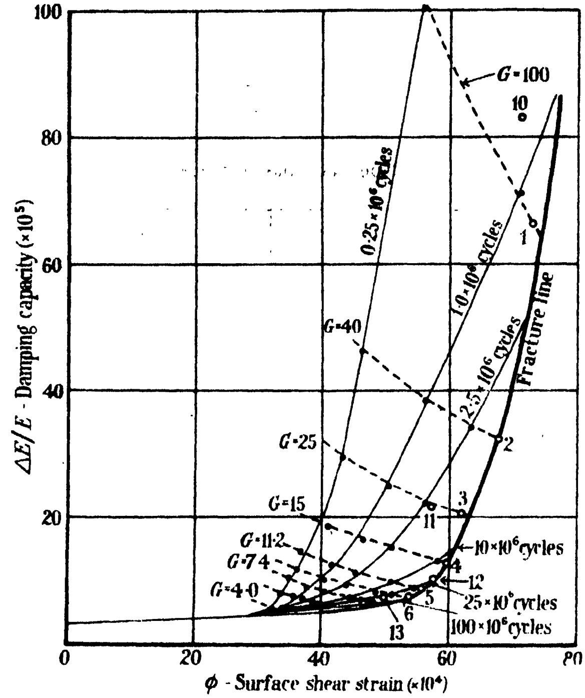


Figure 6. Variation of damping capacity with strain for an alloy of aluminium containing 11% Mg.

It is evident that changes occur very rapidly in the material during the first million cycles, especially for high values of the applied couple (G). The bold curve connects points corresponding to the occurrence of fracture in specimens Nos. 1 to 6.

Alloy containing 5% *magnesium*

The general behaviour of the alloy containing 5% magnesium is similar to that of the 11% alloy and it is not proposed to record detailed observations as in figures 3 and 4. From data similar to those given in these figures, the variation of damping capacity with strain for 0·25 million cycles and at fatigue failure has been derived and is shown in figure 7 together with the corresponding graphs taken from figure 6 for the 11% alloy.

It is evident that the strains which will cause fatigue failure in the 5% alloy are appreciably less than the strains necessary to cause failure of the 11% alloy, the shaded regions in figure 6 showing the limiting conditions of damping capacity and vibrational strain within which fatigue failures have been observed. The figures adjacent to these regions indicate the number of cycles required to cause fatigue failure under conditions of constant peak value of the applied couple. Comparison of these "fatigue regions" with the corresponding curves representing the alloys after vibration for 0·25 million cycles shows that pronounced changes in physical condition occur before failure by fatigue.

Other alloys

Although this paper deals mainly with observations on two alloys, it may not be inappropriate to record that the method of test is applicable to some other aluminium alloys. Figure 8 shows specimens of several alloys in which fatigue cracks have been produced, although no quantitative examination of their fatigue properties has yet been attempted by this method. The specimens shown in figure 8 had natural frequencies of about 1500 cycles per second, and after the completion of the test were anodized to show the fatigue cracks.

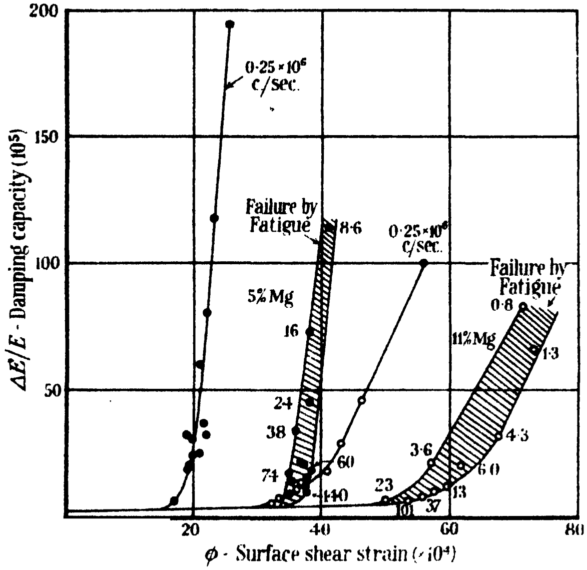


Figure 7. Variation of damping capacity with strain for alloys of aluminium containing 5% and 11% Mg.

§7. DISCUSSION OF RESULTS

For a particular deformation, the damping capacity of the material may, with certain reservations, be regarded as a measure of the ratio of the total plastic strain per cycle to the maximum elastic strain.*

---

* Strictly this may be assumed, for torsional deformation of a solid cylinder, only when the damping capacity is independent of strain. Under such conditions $\Delta E/E = 2\Delta\phi/\phi$, where $\Delta\phi$ is the plastic component of the total strain ϕ. When the damping capacity is not independent of strain, the measured value of the energy dissipated corresponds to an integration of the energies dissipated at strains progressively increasing from zero at the vibration axis to the measured value of ϕ at the surface. For the purpose of discussion it is proposed to assume that $\Delta\phi$ as defined in the above equation represents an average plastic strain for a maximum surface strain ϕ.

Experimental evidence suggests that all materials have a finite, but often very low, damping capacity at very low strains; for the alloys under investigation this small and practically constant damping capacity is estimated to be 3×10^{-5} for surface shear strains up to 30×10^{-4} (5·4 tons per sq. in.) for the 11% alloy and 15×10^{-4} (2·7 tons per sq. in) for the 5% alloy. The occurrence of a small damping capacity at even the lowest strains does not necessarily indicate that purely elastic strains are non-existent; C. Zener (1940) has shown that a small damping capacity may arise from causes other than plastic deformation, although under the conditions of the present investigation these phenomena are not likely to contribute to the observed damping capacity. It is possible that the practically constant damping capacity for small strains may be due to localized slip of a type which may be considered as reversible in the sense that no damage is caused to the crystal structure. The centres of dissipation may be imperfections of the type suggested by G. I. Taylor (1934) and others, where small groups of atoms may be able to assume alternative positions without materially modifying the general structure. Whatever may be the mechanism causing damping capacity at small strains, it is evident that in the range where the damping capacity is small and constant, the alloy shows proportionality between stress and strain.

As the strain is increased above a critical value the damping capacity (for a small number of cycles) increases rapidly (figures 6 and 7) and consequently proportionality between stress and strain no longer occurs; strains of 15×10^{-4} and 30×10^{-4} are, therefore, the limits of proportionality for the 5% and 11% alloys respectively under dynamic conditions which do not involve prolonged vibration. Above this limit of proportionality it is suggested that the mechanism of the dissipation of energy is true plastic strain, i.e. strain which changes the original crystal structure irreversibly. The irreversibility of the process is demonstrated by the dependence of the relations (figures 6 and 7) between damping capacity and strain on the amount of vibration to which the alloys are subjected.

Consequently, at strains above the critical value, the metal may be expected to change in properties during vibration if a proportion of the converted vibrational energy is absorbed as strain energy within the crystal structure. Absorption of energy would result in strain hardening of the metal, and this expectation is confirmed, for figures 4, 6 and 7 show that the damping capacity decreases during continued vibration at strains above the limit of proportionality. Under suitable conditions of vibration the limit of proportionality may be increased from 15×10^{-4} to 34×10^{-4} (2·7 to 6·1 tons per sq. in. shear stress) for the 5% alloy and from 30×10^{-4} to 50×10^{-4} (5·4 to 9·0 tons per sq. in.) for the 11% alloy.

Since strains above the primary limit of proportionality cause changes in the metal it is to be expected that this will be a limiting strain below which fatigue failures will not occur. Figure 7 shows, in fact, that all specimens which failed by fatigue did so at strains greater than the secondary limit of proportionality, i.e. the limit of proportionality after prolonged vibration. The values of damping capacity and strain at which the fatigue failures occurred all lie for each alloy upon a curve which is very similar in shape to that for the damping capacity and strain after a small amount of vibration, but displaced in the direction of higher strain.

It is proposed now to consider the factors which may determine the number of cycles required to cause a fatigue failure. It has been indicated that failure of these alloys is preceded by strain hardening. If $\Delta\phi$ is taken* to represent the amount of plastic strain per cycle (including also the slip corresponding to the damping capacity in the strain range where the damping capacity is constant but small), an expression for the total equivalent plastic strain at fracture is $\Sigma_0^N \Delta\phi$. It is suggested that fatigue failure may depend on this quantity. Another possibility is that fatigue failure is related to the total amount of energy dissipated up to the time of fracture. The total energy dissipated (excluding external losses) is given in table 1 for several specimens of the 11% alloy. Referring at present particularly to this alloy, figures 9*a* and 9*b* show how $\Sigma_0^N \Delta\phi$ and $\Sigma_0^N \Delta E$ vary with N, the total number of cycles to cause fracture (or 10^8 cycles if fracture does not occur before then). The gradient of the line OP represents the average plastic strain per cycle whilst the gradient of the line OQ represents the average dissipation of energy per cycle for specimen No. 7, i.e. the specimen subjected to the highest value of the alternating couple among those specimens which were not fractured after 10^8 cycles. It will be seen that the points corresponding to the fractured specimens (Nos. 1 to 6) lie approximately on a straight line above and roughly parallel to OP or OQ, whilst the points for specimens 8 and 9 lie below these lines.

The criterion for failure by fatigue may be either, from figure 9*a*,

$$\Sigma_0^N \Delta\phi = A + BN \qquad \ldots\ldots(3)$$

or, from figure 9*b*,

$$\Sigma_0^N \Delta E = C + DN, \qquad \ldots\ldots(4)$$

where A, B, C and D are constants. The scatter† of the points representing fatigue failures makes it difficult to decide which of these equations is most likely to represent the truth, but the scatter is somewhat less in figure 9*a* than in figure 9*b*.

In equation (3), the constant B is the amount of slip per cycle which occurs without contributing to fatigue failure of the specimen, i.e. it represents that part of the mechanism causing damping capacity, which exists at all strains and is predominant before true plastic strain occurs. The constant A is the total amount of irreversible strain occurring before failure by fatigue. The constants C and D may be interpreted similarly in terms of energy.

Thus it appears probable that fatigue failures occur either after a certain amount of irreversible strain or, perhaps less definitely, after a certain conversion of energy associated with plastic strain.

It has been stated already that the experimental results suggest that failure occurs after a definite amount of strain hardening, and this would be expected to occur after a definite amount of irreversible strain, but not necessarily after a definite conversion of vibrational energy, since a considerable proportion of the energy represented by the damping capacity may appear as thermal energy. The

---

* It should be noted that under the conditions of operation of the vibration equipment (i.e. constant exciter current) $\Delta\phi$ is practically constant (it would be strictly constant if no external loss of energy occurred) throughout the experiment and is proportional to G, the peak value of the applied alternating couple. It is therefore independent of the amount of strain hardening which may occur during the test.

† This scatter is to be expected in fatigue tests involving a number of specimens because the formation of fatigue cracks is influenced by the conditions of the surface of the metal. The points representing fatigue failure of specimens 10 to 14 of this alloy are also included in figures 9*a* and 9*b*, and it is seen that these also lie above the lines OP and OQ.

energy absorbed as strain energy in the crystal structure is probably roughly proportional to the conversion of energy associated with plastic strain.

The experimental data for the 5% alloy may be interpreted similarly to that of the 11% alloy and it is only necessary to indicate differences in the magnitude of the effects which occur.

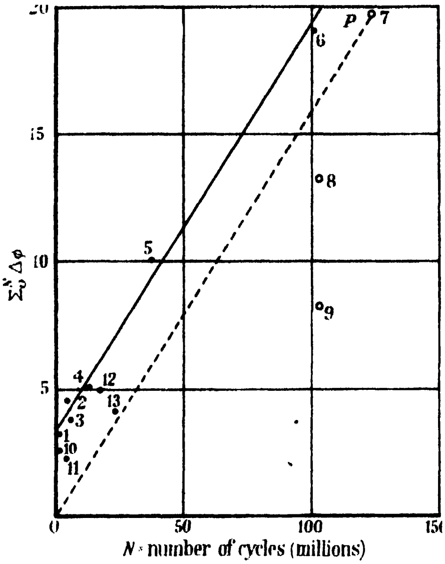


Figure 9*a*. Total "plastic" strain before fracture of aluminium alloy containing 11% Mg.

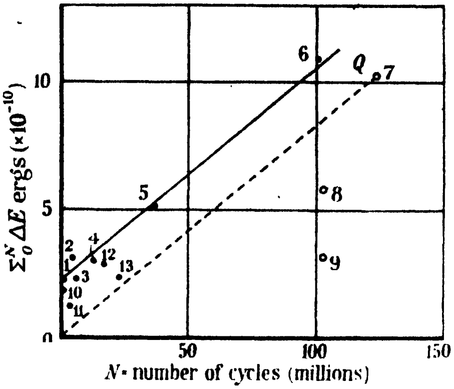


Figure 9*b*. Total energy converted before fracture of aluminium alloy containing 11% Mg.

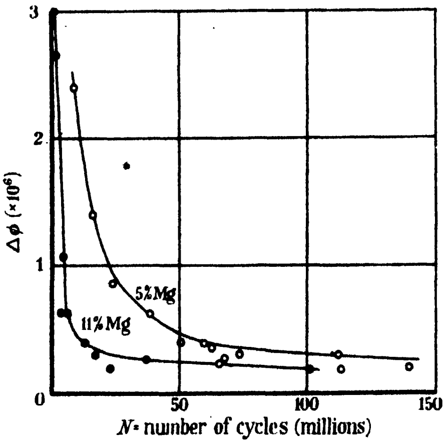


Figure 10. Variation of endurance with plastic strain per cycle.

It is apparent that the endurance of these alloys is dependent on the amount of plastic strain per cycle and the difference between the two alloys is well illustrated by graphs showing endurance as a function of $\Delta\phi$, the plastic strain per cycle, this quantity remaining constant for any one specimen under the present conditions of test. The relation between endurance and $\Delta\phi$ is shown in figure 10 for the two alloys. The curves are roughly hyperbolic, in agreement with equation (3), and indicate that the constant B (slip per cycle not contributing to fatigue failure) is probably the same for both alloys and of the order of $1 \cdot 5 \times 10^{-7}$. The constant A, which represents the total amount of irreversible strain developed before failure by fatigue, is estimated to be 5·8 times as great for the 5% alloy as for the 11% alloy.

§8. CONCLUSIONS

Fatigue failure of aluminium alloys containing 5% and 11% of magnesium is preceded by strain hardening. Strain hardening commences when the shear strain exceeds the primary limit of proportionality of the alloy. Below this strain the damping capacity, which is small (3×10^{-5}) and substantially constant, is not associated with plastic strain as normally understood, but, above this strain, work done in causing plastic strain provides an increasingly important contribution to the damping capacity.

For strains exceeding the primary limit of proportionality, the damping capacity is a function of time of vibration, decreasing as the metal strain hardens to an amount ultimately equivalent to an increase of the limit of proportionality by about 120% for the 5% alloy and about 65% for the 11% alloy. The results suggest that fatigue failures occur when the total amount of irreversible strain caused by vibration reaches a certain value characteristic of the alloy, endurance being dependent on the amount of irreversible strain per cycle.

§9. ACKNOWLEDGMENT

The author wishes to thank the Directors of High Duty Alloys Ltd. for permission to publish this paper.

REFERENCES

Hanstock, R. F. and Murray, J., 1946. *J. Inst. Met.*, **72**, 97.

Taylor, G. I., 1934. *Proc. Roy. Soc.*, A, **145**, 362.

Zener, C., 1940. *Proc. Phys. Soc.*, **52**, 152.

REFRACTION EFFECTS IN ELECTRON DIFFRACTION

By J. M. COWLEY and A. L. G. REES,

Division of Industrial Chemistry, Council for
Scientific and Industrial Research, Melbourne

MS. received 8 *October* 1946

ABSTRACT. Electron diffraction rings from MgO and CdO smokes have previously been shown under high resolution to be multiple. More detailed observations of this phenomenon have been made. It has been observed that the arcs given by tilting of oriented specimens are divided into several components, which vary in intensity and separation with angle of tilt of the specimen. Spotty rings given by brown CdO (containing excess metal atoms) and some MgO specimens are composed of short streaks forming groups of up to six radiating from the expected position of a single-crystal reflection, and in some cases these streaks are divided into inner and outer components.

The theory that the fine structure of the rings arises from refraction at the faces of the regularly shaped particles (cubes) was suggested by Sturkey and Frevel (1945). On the basis of this theory, general expressions have been derived for the deviation of the direction of the diffracted beam owing to refraction at the faces of cubes. Application of these expressions to particular cases has shown complete agreement between the calculated and observed form of the fine structure of the rings, arcs and groups of streaks. Quantitative agreement for the separation of the arc components is obtained by assuming values of inner potential of the order to be expected. The value of the inner potential varies with the reflecting plane from 12 to 16 volts. For the doubled streaks the elongation of the inner component represents a variation of inner potential about this value, arising from excess metal atoms in the lattice. The outer component represents an unexplained value of about 25 volts.

Streaks observed on ZnO smoke rings are explained as due to refraction by the long thin spines characteristic of this material. It is shown that refraction will have the effect of broadening diffraction rings, even for irregular particles, and will give rise to line breadth and intensity anomalies in diffraction patterns from specimens with regular crystal habit.

§ 1. INTRODUCTION

THE development of high resolution electron-diffraction cameras has resulted in the observation of fine structure not previously suspected in electron-diffraction patterns. Hillier and Baker (1945), for example, noted a multiplicity of the rings in transmission patterns from MgO smoke particles; they found the rings 111, 222, 224 and 226 to be double, 200 single, and the 220 ring to possess an unusual contour, suggesting five components. Sturkey and Frevel (1945), in similar experiments on MgO and CdO, made the generalization that *hhh* reflections were invariably double, *h*00 single and others broadened. Sturkey and Frevel attributed this multiplicity to refraction by particles of regular geometrical shape (electron micrographs show them to be perfect cubes), arising from the inner potential (mean potential above free space) within the crystal. This appeared to be confirmed by the variation in the intensity of the reflection from 220 planes observed with change of accelerating potential. By use of a geometrical device referred to as a "triangle of reflection" formed by the intercept of the plane of reflection on the reflecting plane and the cube faces, they derived an expression relating the angular displacement of the components δ to the angles of refraction r_1 and r_2 at the sides of this triangle, viz.:

$$\delta = \frac{P}{2E}(\pm \tan r_1 \pm \tan r_2),$$

where P is the inner potential in volts and E the electron accelerating potential. From this expression P was calculated to be 12 ± 4 volts for the 111 doublet. It will be shown in § 4 of this paper that this expression is a special case of a more general expression and is not generally valid.

In the work described here (see also, Cowley and Rees (1946)) more detailed observations of this phenomenon have been made and several new features of fine structure have been discovered. The theory of refraction by particles of regular shape has been developed fully and has been shown to account quantitatively for the observed features of these patterns. Previously, refraction effects have been considered negligible at the accelerating potentials employed (50 kv.) for all but small reflection angles, but they assume a greater importance in high resolution electron diffraction. The refraction effect is shown to make large contributions to line breadth and, except for particles of diameter <300 A., it represents the predominating factor. Moreover, the unexplained anomalous intensity distributions in certain electron-diffraction patterns are explained quite naturally on the basis of refraction by particles approaching regular geometric forms.

§ 2. EXPERIMENTAL

The electron diffraction patterns were taken with an R.C.A. type EMU electron microscope used with the diffraction adaptor. The precise electron-optical system and the very stable H.T. and lens supplies endow the instrument with high resolution characteristics. Resolution in focused patterns was found to be 12μ or better. Since the system employs an electromagnetic lens between the specimen and the photographic plate to focus the pattern, calibration is necessary to obtain the effective specimen-plate distance. Pure, recrystallized sodium chloride was employed as a standard.

Examination and measurement of fine structural details were made on enlargements to 50 diameters, attained in two stages. Some of the enlargements on which these measurements were made are reproduced in the plate.

The materials studied, namely, MgO, CdO and ZnO, were mainly prepared by the oxidation of the corresponding metal vapour in air. The oxide smoke was collected either directly on a 200-mesh stainless steel gauze or on a thin collodion film (<200 A. thick) supported across the gauze. Materials other than smokes were prepared for examination by the evaporation of suspensions in water or secondary butyl alcohol directly on the collodion film. CdO formed at some distance from the evaporating metal was the usual yellow-orange colour ; that formed close to the metal was dark brown, since conditions are here more favourable for incomplete oxidation with the formation of a non-stoichiometric oxide containing an excess of cadmium in the oxide phase. As it happens, these two forms of oxide exhibit significant differences in the fine structure of the diffraction patterns which will be discussed later.

Data on particle size and shape were obtained from electron micrographs. Both MgO and CdO smokes consist of perfect cubic particles, and ZnO smoke of fine needles arranged in "fourlings" together with occasional semi-transparent plates. For MgO and yellow CdO the average cube edge was ~500 A. ; for brown CdO the particles were somewhat larger (cube edge ~1000 A.) and there were present some sintered aggregates of small numbers of particles. In ZnO smoke, it is shown that the [0001] axis is parallel to the axis of the spikes and it is probable, although difficult to establish by direct observation, that the spikes have hexagonal cross-section.

§ 3. RESULTS

The results will be discussed in relation to three types of diffraction pattern observed, viz. :

(i) Continuous ring patterns showing subdivision of the normal reflections.

(ii) Arc patterns obtained from specimens possessing a preferred orientation.

(iii) Ring patterns showing the individual single crystal reflections as groups of spots or streaks, which occurred when only a small number of crystals were present in the irradiated area of the specimen.

Continuous ring patterns

Patterns given by MgO and CdO smokes showed a subdivision similar to that previously reported. Our observations, tabulated in table 1, are more detailed than those referred to.

Table 1

Reflection	Multiplicity	Remarks
*h*00	Single	Sharp
hhh	Doublet	Clearly resolved
*hh*0	Triplet	Central component sometimes very weak, resembling doublet
*hk*0	Triplet	——
331 422 531	Doublet	——
331, 511, 442, 622, 642	Triplet	——

The two components of a doublet appear one on either side of the normal position. In triplets, the central component is undisplaced and the outer components displaced symmetrically. In figure 1 (*a*) an enlargement of a portion of one of these patterns showing resolution of these components is given.

Arc patterns

In many specimens supported on collodion films, a large majority of the cubes were oriented with the cube face parallel to the film, so that, on tilting the film away from the position normal to the beam, a pattern of arcs lying on layer-lines corresponding to $l = 0, \pm 1, \pm 2$ was observed for *hkl* planes recorded. These arcs showed multiplicity and, in view of the greater intensity in the arcs, more information could be obtained regarding the number and separation of components, since the weaker components were more evident. The intensities and separations of the components varied with the angle of tilt and were, in fact, different for the different arcs on the one ring.

Component separations, measured on a travelling microscope, and visually estimated intensities, are given for both rings and arcs for CdO in table 2. A selection of the data only is recorded, since the complete set of data would be too extensive to reproduce here. Data have been obtained for arcs on CdO patterns at seven angles of tilt between 0° and 60°. The components are represented by letters indicating their relative intensity (strong (S), medium (M) and weak (W)) and the letters are separated by numbers indicating the separation of these components in units of 10^{-5} radian. Thus the record, W42S44S42W, indicates that two strong components occur at positions $\pm 22 \times 10^{-5}$ rad. from the expected position and that two weaker components occur outside the strong component at $\pm 64 \times 10^{5}$ rad. displacement. It will be noted that there is a marked change in the intensities and separations of components of any one arc with increased

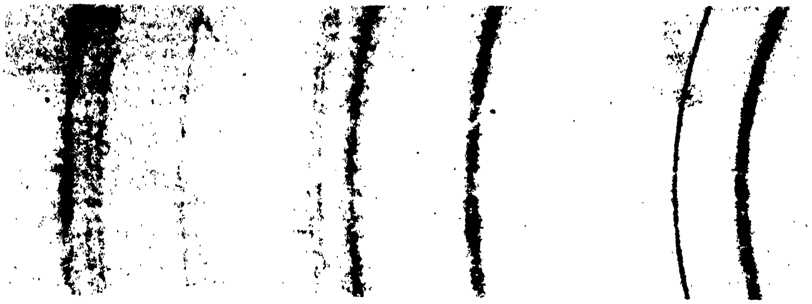


Figure 1 (*a*). Electron diffraction pattern from yellow cadmium oxide showing resolution of ring components. Enlargement 16 diameters.

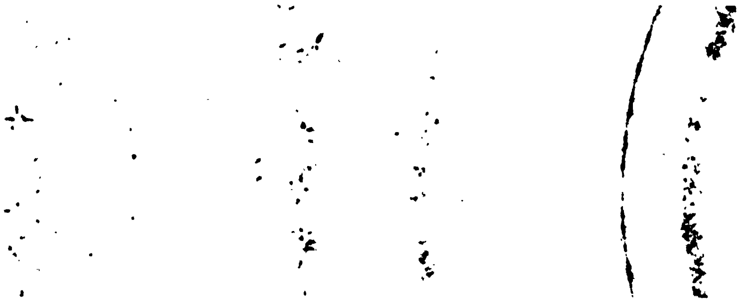


Figure 1 (*b*). Electron diffraction pattern from brown cadmium oxide (excess cadmium) showing resolution of single crystal reflections. Enlargement 16 diameters.

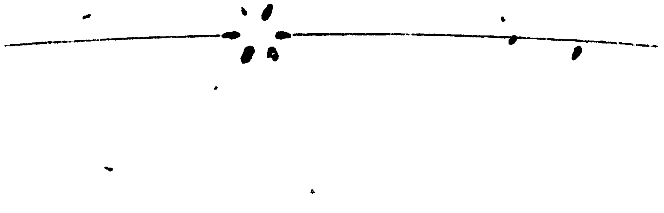


Figure 2. Group of six streaks on 422 ring arising from single crystal of magnesium oxide. Enlargement 56 diameters. Undisplaced ring position indicated by continuous line.

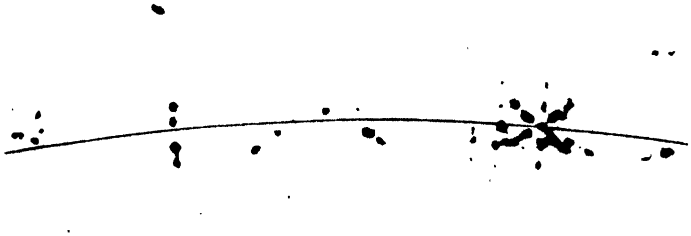


Figure 3. Groups of streaks on 220 ring arising from single crystals of magnesium oxide. The second component on each streak is clearly visible. Enlargement 56 diameters. Undisplaced ring position indicated by continuous line.

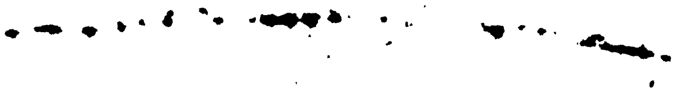


Figure 4. Single crystal reflections along the 200 ring. Enlargement 56 diameters.

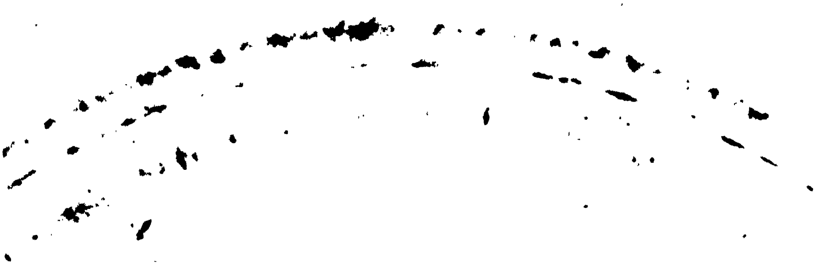


Figure 12. Portion of diffraction patterns from zinc oxide smoke particles showing streaks on 100, 002, 101 rings. Enlargement 24 diameters.

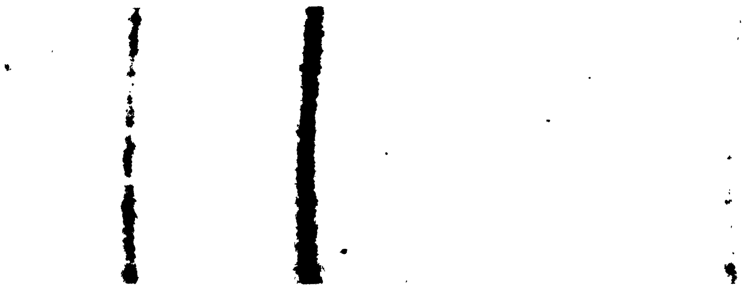


Figure 13. Portion of diffraction patterns showing 222, 400, 420 and 422 rings from near cubic particles. Enlargement 24 diameters.

angle of tilt. Examples of this behaviour are shown in figures 7, 8 and 9, in which the theoretical relation between separation and relative intensity and angle of tilt are compared with the experimental results.

Patterns showing single crystal reflections

Diffraction patterns from specimens having a relatively small number of individual crystals in the illuminated area exhibit "spotty" rings and the spots appear to be grouped together in groups of two to six, where the spots lie on the various ring components. This grouping was more easily observed in patterns from brown CdO and MgO collected close to the metal surface, in which the spots were elongated into short streaks radiating for each group from a common point lying on the normal ring position (figure 1 (*b*)). Groups of up to six streaks were observed with the undisplaced component of a group of spots usually absent. This phenomenon is illustrated in figures 2 and 3. There is a noticeable tendency for these streaks to make angles with the radius of the pattern which are characteristic of the ring in question. For example, for both MgO and CdO the angles are (i) 45° for 111, (ii) 0° or 45° for 220, (iii) 90° for 200, etc. For the 200 reflection, the two components lie along the ring, which therefore appears single, a fact which explains the relative sharpness of this ring. This is illustrated in figure 4. For patterns showing streaks and groups of streaks, the angles between the streaks and the radius of the ring were measured for various rings on 50-diameter enlargements. Definite maxima appeared in the frequency plots, confirming the existence of preferred angles. Maxima were found in such plots for CdO at 30° to 40° for the 222 ring, at 0° and 50° for 220 and at 15°, 30° and 70° for 311. One of the plots is reproduced in figure 10 accompanying the theoretical discussion of these results in § 4.

Another feature of certain of these patterns, particularly evident in the inner rings, is the presence of a second component of each streak in a group; this is illustrated in figure 3. It was found that the ratio of distances from the centre of a group to the outer and inner components was constant for a given ring. This observation is discussed in § 7.

§ 4. THEORETICAL

Many phenomena observed in x-ray and electron diffraction, which bear some resemblance to the effects described, suggest interpretations different from that developed later in § 4. It was considered desirable to examine these effects and to show where these interpretations failed.

In many single crystal x-ray patterns, "extra" spots, usually temperature dependent and diffuse, are observed and these have been satisfactorily explained by the Faxén–Waller theory as arising from thermal vibration of the atoms comprising the crystal. Most reminiscent of the effects described in § 3 are the secondary reflections observed by Lonsdale (1945) in type 1 diamonds; these spots are sharp and not markedly temperature sensitive. W. H. Bragg (1942) has interpreted these extra spots on the basis of extended reciprocal lattice points in small crystals. Similarly, v. Laue (1936) has explained the occurrence of groups of spots in electron-diffraction patterns as arising from the shape of the crystals. He pointed out that each reciprocal lattice point would possess horns or extensions

Table 2

Ring	Ring profile	l	$\alpha=10°$	27°	29°	40°	46°	52°	60°
111	W51S41S51W	1				W42S44S42W	S41S	S24S	W46S41S46W
220	W44S24S44W	0	W61S61W	S44M44S	S60S	S43S	W55S46S55W	S37S	W49S35S49W
		2					W47S31S47W	W46S34S46W	W57S52S57W
311	W44S44W	1		S41S	W39S29S39W	W49S49W	W46S46W	W43S43W	W44S21S44W
420	W42S42W	0	W46S46W	S44S	W51S32S51W	W46M28M46W	W44M22M44W	W41M19M41W	W35S17S35W
		2		M46M46M	W44M44W	W43M43W	W38M20M38W	M23M	M17M
440	—	0	W62M62W	M41W41M	M62M	W60W	W62W	W51W	W43W

normal to the bounding faces of crystals and that these would be more pronounced for small crystals of regular shape. On this basis v. Laue was able to explain the results of Cochrane (1936) and Brück (1936) for layers of Ni, Co and Ag on various substrates.

Whilst these effects superficially resembled those reported here and the order of crystal size involved in both cases was the same, they differed in several significant respects. All these apparently related effects are of a different order of magnitude, and in every case the central component of observed groups of spots would necessarily be strong, whereas in the single crystal groups observed in our patterns the central spot is always absent. Moreover, on v. Laue's theory, the extensions associated with each reciprocal lattice point for cubes would be symmetrical and independent of orientation of the crystal, whereas the structures of arcs on the one ring have been shown to differ widely and also to vary with orientation. Predictions based on this theory regarding the subdivision of rings, the distribution of angles between streaks and radius, and the angles between streaks and radius in individual groups of spots, did not agree with the observations recorded in § 3.

Daniel and Lipson (1943) observed side bands in X-ray powder patterns from the phase Cu_4FeNi_3, attributable to the presence of a superlattice formed by the regular segregation of different atoms. However, a strong central component is expected here also. An interpretation on this basis would require *h*00 rings to be doubled, whereas our observations show them to be single.

Rooksby (1943) has reported similar multiplicity in X-ray powder patterns from NiO. *h*00 rings were observed to be single, whilst other reflections were either double or triple. Rooksby was able to account for these observations by assuming that the structure was only pseudo-cubic and was, in fact, rhombohedral. This interpretation cannot be used for the results of this investigation since it does not account for the grouping of spots or streaks; moreover, spots would be undivided and on one component or the other.

X-ray powder patterns of the MgO and CdO used, taken with a 14 cm. camera, exhibited no anomalies of the type found in electron diffraction.

Refraction theory

The explanation of the fine structure features on the basis of reflection suggested by Sturkey and Frevel (1945) has been investigated more thoroughly and found to predict the observed effects accurately. The vector method of Luneberg (1944) has been applied to the general case.

For reflection or refraction at a plane surface L_i (figure 5),

$$\overline{S}_{i+1} = \overline{S}_i + \Gamma_i \,.\, \overline{M}_i.$$

$\overline{S}_i = n_i \overline{T}_i$, where n_i is the refractive index and $\overline{T}_i$ the unit vector along the beam. Γ_i is a scalar function of the angle of incidence ϕ_i and $\overline{M}_i$ is the vector perpendicular to the plane L_i. Putting $\rho_i = n_i \cos\phi_i = \overline{S}_i \,.\, \overline{M}_i$, we have $\Gamma_i(\rho_i) = -2\rho_i$ for reflection and $\Gamma_i(\rho_i) = (n_{i+1}^2 - n_i^2 + \rho_i^2)^{\frac{1}{2}} - \rho_i$ for refraction.

For a ray passing through p faces

$$\overline{S}_{p+1} = \overline{S}_1 + \sum_{i=1}^{p} \Gamma_i(\rho_i) \,.\, \overline{M}_i.$$

For a ray passing through two faces of a cube, with normals $\overline{M}_1$, $\overline{M}_2$, and being reflected internally at the Bragg angle θ at a plane with normal $\overline{N}$ (figure 6), we may write

$$\overline{S}_4 = \overline{S}_1 + \Gamma_1(\rho_1)\overline{M}_1 + \Gamma_N(\rho_N)\overline{N} + \Gamma_2(\rho_2)\overline{M}_2,$$

which may be reduced to

$$\overline{S}_4 = \overline{S}_1 - 2n\sin\theta\,.\,\overline{N} + \frac{n^2-1}{2\cos\phi_1}\,.\,\overline{M}_1 - \frac{n^2-1}{2n\cos\phi_2}\,.\,\overline{M}_2.$$

Since $n^2-1 = P/E$ for an inner potential of P volts and an accelerating voltage of the beam E, and n is close to unity, we have, taking account of the possible variation in signs,

$$\overline{S}_4 = \overline{S}_1 - 2\sin\theta\,.\,\overline{N} + \frac{P}{2E}\left(\pm\frac{\overline{M}_1}{\cos\phi_1} \pm \frac{\overline{M}_2}{\cos\phi_2}\right).$$

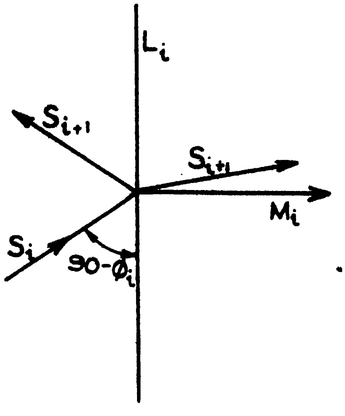


Figure 5.

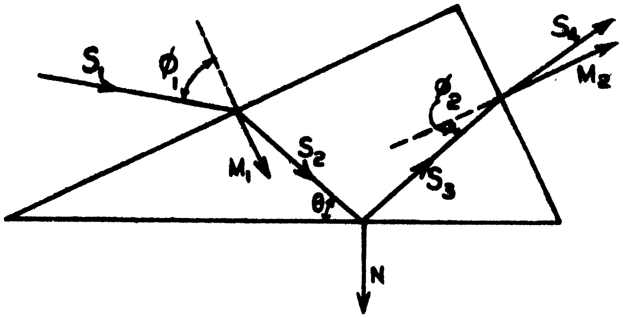


Figure 6.

Here the second term represents the deviation due to Bragg reflection, and the third term represents the added deviation due to refraction. The direction of the refraction displacement will be governed by the direction of the surface normals. For cubes there are, in general, six pairs of non-parallel faces through which the beam may pass, corresponding to various combinations of the face normals $\pm\overline{M}_1$, $\pm\overline{M}_2$, $\pm\overline{M}_3$. Hence there may be deviations in six directions, giving rise to a group of up to six spots surrounding the normal spot position, or, in the case of varying inner potential, six streaks radiating from the common centre. Pairs of parallel faces give no displacement. The undisplaced spot is usually weak and often missing owing to absorption in passing through the whole length of the cube. This theory then gives at least a qualitative explanation of the groups of spots and streaks.

Arc patterns

In the patterns of arcs given by oriented cubic crystals on tilted films, each arc is given by many cubes in almost the same orientation with respect to the beam, and hence the arc will be generated by translating a certain spot-group along the circumference of the ring. Thus the arcs will be multiple and the distribution of components across the arc will be given by the projection of the appropriate group of spots on the radius, that is, by projecting the deviations δ upon the direction of N, the normal to the diffracting plane.

Suppose the beam makes angles ϕ_1, ϕ_2, ϕ_3 with the normals $\overline{M}_1$, $\overline{M}_2$, $\overline{M}_3$, and these normals make angles ψ_1, ψ_2, ψ_3 with $\overline{N}$. Assuming the Bragg angle θ and the angular deviation δ to be negligibly small compared with ϕ_1, ϕ_2 and ϕ_3, the projections of the deviations δ on $\overline{N}$ are of the form

$$\delta_N = \frac{P}{2E}\left(\pm\frac{\cos\psi_1}{\cos\phi_1} \pm \frac{\cos\psi_2}{\cos\phi_2}\right).$$

In the special case where the path of the beam lies in a plane perpendicular to a cube edge, this becomes

$$\delta_N = \frac{P}{2E}\,(\pm\tan\phi_1 \pm \tan\phi_2),$$

which is the form given by Sturkey and Frevel (1945).

In general, this degeneration is not valid, since the path of the beam is not normal to the cube edge.

In an arc there will be a maximum of seven components including the component with zero displacement; equal indices or zero indices of the diffracting plane will reduce this number. The relative intensities of the components of an arc may be calculated from consideration of the areas presented to the beam corresponding to each component. Two limiting cases may be distinguished—(i) if absorption is neglected the whole area of a cube presented to the beam is effective; (ii) for very high absorption only a thin strip along the cube edges will contribute and the width of the strip will vary with the angles involved.

For the (220) *plane*, for example, the angles between face normals and $\overline{N}$ are $\psi_1 = 90°$, $\psi_2 = \psi_3 = 45°$ and

$$\phi_2 = \phi_3; \qquad \cos\phi_2 = 1/\sqrt{2}\sin\phi_1.$$

The possible values of $(\delta_N \,.\, 2E/P)$ will then be

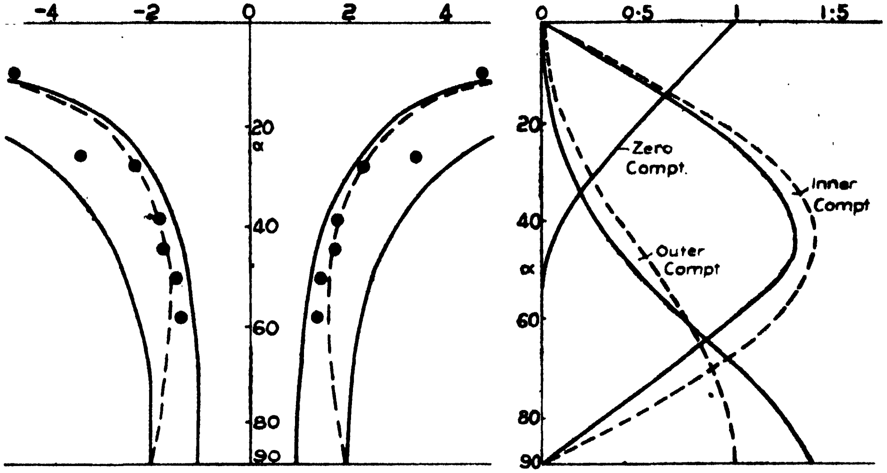
$$0, \quad \pm\frac{1}{\sin\phi_1}, \quad \pm\frac{2}{\sin\phi_2}.$$

Then ϕ_1 will be the angle of tilt α of the specimen with respect to the beam, and so the displacement of the components of the 220 arc may be plotted against α to give the curves of figure 7 (*a*). The relative intensities of the components are calculated readily for the two limiting cases of zero and high absorption. With the exception that in the latter case the zero displacement component is absent, there is no great difference in the relative intensities, as is shown in figure 7 (*b*). In agreement with observations (table 2) the zero displacement component decreases in intensity with angle of tilt. In practice the two displaced components are not resolved. The dotted curve of figure 7 (*a*) represents a mean displacement weighted for intensities. This is plotted for the case of no absorption, which differs very little from that for high absorption.

The measured displacements, δ_N, from CdO patterns for various angles of tilt are plotted as full circles in figure 7 (*a*), with a multiplying factor of $2E/P$, where P is chosen as 12·7 volts to give the best agreement. The value of the inner potential for CdO on the assumption of no absorption is thus 12·7 ± 2·0 volts, and for high absorption it is 12·3 ± 2·0 volts.

Measured values of the displacements for the 440 arc fitted the same curve for $P = 15 \pm 2$ volts.

Corresponding calculations for the 202 arc on the $l = 2$ layer line are made by taking the angle of tilt $\alpha = \phi_2$. This arc does not appear until $\alpha = 45°$, as is shown in the graph of displacement against α in figure 8. The relative intensities are indicated approximately by the thickness of the curves. Similarly, the displacements for the $l = 0$ and $l = 2$ arcs on the 420 ring have been plotted in figure 9.



(*a*) (*b*)

Figure 7 (*a*). Plot of calculated displacement of the components of the 220 arc against angle of tilt. Experimental data—full circles.

(*b*). Plot of relative intensity of components of the 220 arc against angle of tilt. Continuous line—zero absorption. Broken line—high absorption.

The outer two components will not be resolved and a curve is therefore drawn representing the mean. The plotted points are for CdO arcs with $P = 16·7$ volts for $l = 0$ and $P = 15·6$ for $l = 2$. Similar calculations have been made for 111 and 620 arcs. Values of inner potential found for the various arcs are tabulated in table 3. Those for which the accuracy is not indicated are less accurate than the others.

Table 3

Arc		CdO	MgO
111	$l=1$	11	13
220	$l=0$	12·7±2·0	—
	$l=2$	10	—
222	$l=2$	14	17
420	$l=0$	16·7±1·5	15
	$l=2$	16	15
440	$l=0$	15·0±2·0	15
	$l=4$	13	10
620	$l=0$	15	11
	$l=2$	17	12

It will be noticed that the values of P for the second orders of 111 and 220 arcs are higher than for first orders. The difference is within the possible experimental error, but is sufficiently consistent to indicate a real difference. The error introduced in the calculations by assuming the Bragg angle θ to be negligible is in the right direction, but further calculations have shown that the magnitude of the error introduced by this assumption is not sufficiently large (less than about 3%, except for very large displacements of very weak components of the arcs) to explain this difference.

The variation of the value of inner potential between different reflections is greater than the probable experimental error, and indicates that there is a dependence of the effective inner potential on the direction of the electron path through the crystal. The values of the inner potentials obtained for CdO and MgO are of the order expected, but, apart from the value 12 ± 4 volts given by Sturkey and Frevel (1945), there are no values given in the literature with which these results may be compared. However, the qualitative and quantitative agreement of the observed forms of the arcs and those calculated on the basis of the refraction theory is complete within experimental error, provided these values of P are taken as correct.

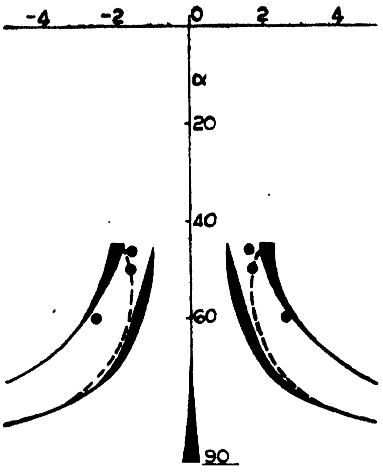


Figure 8. Plot of calculated displacement of components of 202 arc against angle of tilt. Breadth of curve gives indication of dependence of relative intensity on tilt. Experimental data—full circles.

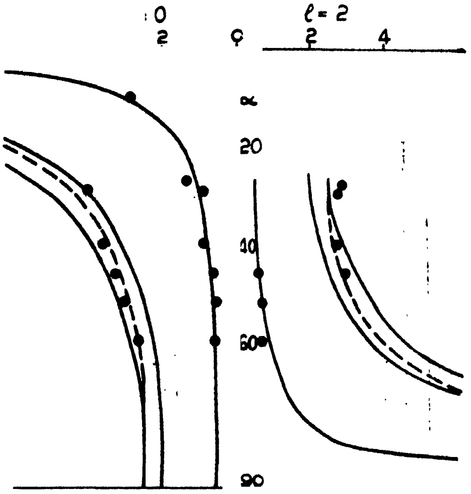


Figure 9. Plot of calculated displacement of components of 420 arc, for both $l=0$ and $l=2$, against angle of tilt. Experimental data—full circles.

Patterns showing streaks and groups of streaks

The form of a group of streaks for any given orientation of a cube giving a particular reflection may be predicted from the refraction theory. Without detailed calculations it is possible to derive general conditions satisfied by the angles between the streaks of a group and the radius for any particular ring. The three pairs of diametrically opposite streaks will define three angles with the radius. For the 200 ring there will be only one pair of streaks and the corresponding angle will be 90°, i.e. the streaks lie along the ring; for 220, one angle will be 0°, the other two will be equal and lie in the range 45–90°; and so on.

Many measurements were made of the angles between the streaks of individual groups and the radius, and in every case the angles complied with the above conditions within the limits of experimental error.

Calculations of the distribution of the angles between streak and radius were made for the 220 ring. The probability of a given angle between streak and radius occurring was calculated on the assumption that all angles of tilt α are equally probable. This probability was multiplied by the intensity of the streak to provide a measure of the chance of that particular angle being observed. This product has been plotted against angle in figure 10. The number of angles measured in ranges of 4° is plotted in the same figure as a block-diagram. Allowing for the inaccuracy of measurement of angles the agreement is satisfactory.

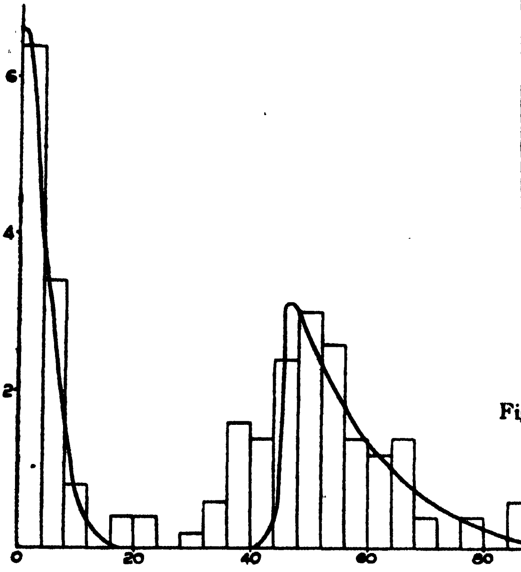


Figure 10. Calculated distribution of angles made by individual streaks with radius of diffraction ring for 220 reflections. Experimental data plotted as block diagram.

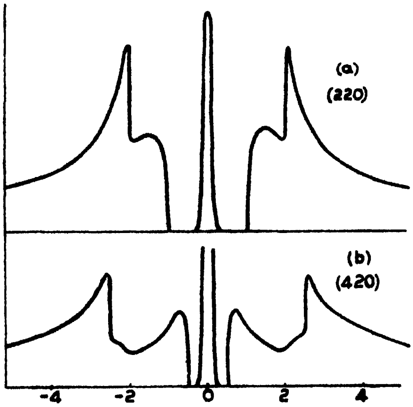


Figure 11. Calculated intensity profile for (*a*) 220 ring, (*b*) 420 ring. Relative intensity of central components will vary markedly with crystal size.

§ 5. RING PROFILES AND RING BREADTH

Cubic crystals

Rings given by random cubes may be considered as the sum of the arc patterns for all angles of tilt in all positions.

From the method outlined in the discussion of arc patterns in § 4 for obtaining the displacement and intensity of each component in terms of the angle of tilt, the intensity of each component may be expressed as a single-valued function of displacement. Summing the intensities due to the several components for each displacement gives the intensity profile of the continuous ring. This has been done for the 220 and 420 rings, and the profiles obtained on the assumption of no absorption are shown in figures 11 (*a*) and 11 (*b*). The profiles calculated for the case of high absorption are closely similar, except for the absence of the central component. In practice the broadening of each part of the profile due to finite crystal size and finite beam diameter will permit the resolution of only the gross features of the contours. Thus the 220 ring is a doublet for large crystals and close triplet for small crystals. The three central components of the 420 ring are not resolved, and the ring then appears as a triplet.

Zinc oxide smoke patterns

The rings from regularly shaped particles other than cubes will similarly have characteristic profiles. In the case of ZnO smoke particles, only the prism faces parallel to the hexagonal axis are developed in the long spines of which this material is composed. Hence the 00*l* rings will be sharp and single, since the streaks due to refraction will lie along the rings; the *hk*0 rings will have streaks perpendicular to the rings, and hence these rings will be broadened; the streaks for *hkl* rings will be inclined at various angles. These points are well illustrated by figure 12, which is an enlargement of part of the inner three rings of a ZnO smoke pattern. From the centre outwards the rings are then 100, 002 and 101, and it will be seen that the inclination of the streaks is as predicted.

Hillier and Baker (1946) have recently published enlargements of ZnO smoke patterns showing similar streaks, which they interpreted as low magnification electron-optical images of the individual ZnO spines. Such images, however, would be radial on the 002 ring and circumferential on the 100 ring, whereas the opposite is the case, so that this interpretation of the streaks cannot be correct.

It is evident from the above that the breadth of the continuous rings in patterns from finer ZnO smoke will be different for different rings; the 00*l*, for instance, will be much sharper than *hk*0 rings.

Irregular particles

Deviation from perfect regularity of, for example, the CdO and MgO cubes will prevent the resolution of the ring components. Selectively broadened rings, in which the components are not resolved (figure 13), were obtained from commercial MgO, shown by the electron microscope to consist of cubes with rounded corners. For completely irregular particles, or the similar case of spherical particles, the refraction effect will broaden all rings by the same amount, as our observations have confirmed. The broadening of the rings for this case may be calculated as follows:—Considering a point reflection due to a single crystal, the intensity at a distance r from the point, in suitable units, will be

$$I_r = \cos\phi_1 \,.\, \cos\phi_2 \,.\, d\phi_1 \,.\, d\phi_2.$$

If r_1 and r_2 are the component displacements due to the two faces traversed, $r_1 = q\tan\phi_1$, where $q = P/2E$, and

$$I_r = q^4(r_1^2+q^2)^{-3/2}(r_2^2+q^2)^{-3/2}\,dr_1 \,.\, dr_2.$$

Taking the radial displacements $p_1 = r_1\cos\gamma_1$ and $p_2 = r_2\cos\gamma_2$, projecting the above distribution on the radius and integrating for all values of γ_1 and γ_2, one obtains for the intensity distribution across the ring

$$I_p = \left(1 - \frac{p_1}{(p_1{}^2 + q^2)^{\frac{1}{2}}}\right)\left(1 - \frac{p_2}{(p_2{}^2 + q^2)^{\frac{1}{2}}}\right),$$

where

$$p = p_1 + p_2.$$

For spherical particles, where $p_1 = p_2$, the half-width of this distribution is 1·4 $P/2E$, and the width for one-tenth intensity (approximately the line breadth estimated visually) is 3·8 $P/2E$. For values $P = 15$ volts and $E = 50$ kv. these widths correspond to 0·6′ and 1·8′ respectively. For particles of irregular shape $p_1 \neq p_2$ and the half-width will be approximately 1·2 $P/2E$.

The above calculation is made for the case of low absorption. For high absorption the central peak of the distribution will be flattened, and the half-width correspondingly greater.

This broadening of the rings due to the refraction effect will be additional to that due to the factors formerly recognized as contributing to ring breadth. These are finite beam width, lack of homogeneity of electron velocities, and finite crystal dimensions. The first two factors may be reduced to negligible proportions by suitable design of apparatus, so that only the last need be considered here. According to Brill (1934) the breadth associated with near-cubic particles of mean cube edge Λ is $0{\cdot}94\lambda \sec\theta/\Lambda$. The broadening due to refraction will be of the same magnitude as that due to finite crystal dimensions for crystal size of the order of 250 A. for $P = 15$ volts and $E = 50$ kv., and the error due to neglect of the refraction effect will still be 10% for 25 A. crystals.

In the case of regularly shaped crystals, the selective broadening of some rings by refraction will be as much as two or three times greater than that for irregular shapes. This will complicate the estimation of crystal shape by comparison of ring breadths. In general, the refraction effect will give differences in the same direction as those due to unequal crystal dimensions. For example, the *hk*0 rings from ZnO smoke will be selectively broadened by refraction and also by the small crystal dimension across the long spines.

§ 6. RELATIVE INTENSITY OF RINGS

For particles of regular shape it will be evident that the peak intensity of a ring which is multiple will be much less than that for an unbroadened ring, although the integrated intensity is identical. In the intermediate case, in which components are not resolved, and for particles of irregular shape, where the ring is simply broadened, a similar situation will obtain. Since ring intensity is usually derived from peak intensities, the intensity distribution will be found to deviate further from those calculated from structure factor for particles tending to assume regular shapes. The unbroadened *h*00 rings for cubes will therefore be relatively strong. Intensity measurements should, of course, always be made by finding the area under ring contours by precise microphotometry. Even so, it may be seen from the ring profiles for 220 and 420 given in figures 11 (*a*) and 11 (*b*) that a considerable proportion of the reflection intensity will, as a result of refraction, appear at relatively large deviations in broadened rings. This part of the intensity will be indistinguishable from background on the photographic plate, with the result that the measured intensities of the broadened rings will be too small by comparison with those of sharp rings.

ZnO smoke gives electron-diffraction patterns which illustrate this effect. Yearian (1935) and others have observed that planes parallel, or nearly so, to the basal plane give reflections that are invariably too strong. Thus reflections from 002, 102 and 103 have intensities greater than theoretical: x-ray diffraction shows the line intensity of 110 > 102 and 103, whereas electron diffraction shows 103 and 102 > 110. This is precisely the way in which refraction for long thin hexagonal prisms would affect the intensities.

It is expected that intensity differences due to refraction will become relatively less important as the crystal size becomes smaller or if resolution is poor. While no quantitative measurements of this effect have been made, the same qualitative intensity distribution has been reported for ZnO smoke examined with cameras of poorer resolving power than that used in this work. Moreover, ZnO specimens prepared by other methods, having less regular particle shape, as shown by electron-microscope examination, give electron-diffraction patterns in which the intensity distribution follows the theoretical closely; in particular, the 110 ring becomes more intense than 102 and 103.

§7. FURTHER NEW FEATURES OF ELECTRON DIFFRACTION PATTERNS

As mentioned in §3 and illustrated in figures 2 and 3, patterns from brown CdO and certain specimens of MgO contain streaks and groups of streaks in place of the expected spots on ring components. The streaks in one group all radiate from a common centre—that is, δ varies in magnitude, but not in direction. The only factor which could account for this type of variation in δ is a variation of the inner potential. Similarly, the second component observed for streaks on the inner rings lies on the same radial line and must therefore be attributed to a second higher value of the inner potential.

In instances where the direction of the streaks represented a recognizable orientation of the crystal so that the crystal faces involved in the refraction could be identified, it was possible to calculate from measurements made in the variation of δ the inner potentials corresponding to various parts of the streaks. In general, the elongation of the inner component of the streaks represented a variation of the inner potential about that found from the arc patterns. For the (220) plane, however, the average value was found to be about 8 volts, somewhat less than the value derived from the 220 arc. The outer component of the streaks corresponds in every case to an inner potential of 25 volts, even for cases in which the potentials corresponding to the inner components differ. The 111 ring of the CdO pattern shows a third component corresponding to an inner potential of 40 to 50 volts.

The variation of the inner potential about the mean value may be interpreted as arising from the presence of excess Cd atoms in inter-lattice sites in the non-stoichiometric oxide. The mean inner potential would change in the region of such defects, and electrons passing through such regions would be refracted accordingly. The presence of components corresponding to inner potentials of 25 volts and more has not yet been explained.

REFERENCES

Bragg, W. H., 1942. *Proc. Roy. Soc.*, A, **179**, 51 and 94.
Brill, R., 1934. *Z. Kristallogr.*, **87**, 275.
Brück, L., 1936. *Ann. Phys., Lpz.*, **26**, 233.
Cochrane, W., 1936. *Proc. Phys. Soc.*, **48**, 723.
Cowley, J. M. and Rees, A. L. G., 1946. *Nature, Lond.*, **158**, 550.
Daniel, V. and Lipson, H., 1943. *Proc. Roy. Soc.*, A, **181**, 368.
Ehrhardt, C. H. and Lark-Horovitz, K., 1940. *Phys. Rev.*, **57**, 603.
Hillier, J. and Baker, R. F., 1945. *Phys. Rev.*, **68**, 98.
Hillier, J. and Baker, R. F., 1946. *J. Appl. Phys.*, **17**, 12.
v. Laue, M., 1936. *Ann. Phys., Lpz.*, **26**, 55.
Lonsdale, K., 1945. *Nature, Lond.*, **155**, 572.
Luneberg, R. K., 1944. *Mathematical Theory of Optics* (Brown Univ.), Appendix II.
Rooksby, H. P., 1943. *Nature, Lond.*, **152**, 304.
Sturkey, L. and Frevel, L. K., 1945. *Phys. Rev.*, **68**, 56.
Yearian, H. J., 1935. *Phys. Rev.*, **48**, 631.

ELECTRON OPTICS AND SPACE CHARGE IN STRIP-CATHODE EMISSION SYSTEMS

By O. KLEMPERER,
Research Laboratories, Electric and Musical Industries, Ltd.,
Hayes, Middlesex

(Now at Imperial College, London)

MS. received 2 September 1946

ABSTRACT. A description of strip cathode systems is given and their merits are compared with those of circular emission systems. It is shown how electron-optical laws can be applied to simple strip systems as long as the emission is sufficiently small. Results obtained under space-charge conditions reveal that the beam spread, due to mutual repulsion of the electrons, completely changes the emission distribution. These results, however, can be interpreted qualitatively by a simple space-charge theory.

Special systems, proposed by Pierce, in which the electron-optical orbits are not upset by space charge, are investigated by measuring the current that can be transmitted through a tunnel of given dimensions. Large discrepancies are found between experimental results and predictions of the simple space-charge theory.

With a view to finding the reasons for these discrepancies, ray-tracing results are discussed and the potential distribution in beams of relatively large space charge is investigated. Tracing results reveal lack of homocentricity, which is an essential supposition in the simple theory of beam spread. The lack of homocentricity, however, is found to be caused by the potential distribution in the beam. This potential distribution is probed in a large beam of high space charge by means of a fine beam of low intensity.

§ 1. INTRODUCTION

STRIP-CATHODE emission systems are electron sources to produce flat, ribbon-shaped or wedge-shaped beams which, by a focusing lens, can be projected into a line focus. The glass-optical counterpart of the strip-cathode system is the elongated light source, the rays of which may be projected into a line focus by means of glass lenses with cylindrical surfaces. The glass-optical and the electron-optical line focus systems can be treated geometrically as two-dimensional problems provided the sources are sufficiently long to render negligible disturbances arising from conditions at the ends. The paths of the rays, however, are entirely different in the two systems, since space-charge effects always play an important part in the electron-emission systems.

Both in light optics and in electron optics the line focus systems have been applied far less frequently than systems of circular symmetry, and for this reason their properties have not received much attention in the past. However, their investigation should be of certain basic interest. Due to the greater simplicity of the two-dimensional geometry, the problems of the strip systems are in many cases more accessible to simple argument. Moreover, a comparison of the different properties of strip systems and circular systems will bring out various interesting points. The voltage-current relations of strip systems and circular systems having the same cross-sectional electrode arrangement are entirely different. For a given cross-section and at a given anode voltage, the strip system will always yield appreciably larger currents, even if the strip cathode is considerably restricted in length. The reason for this is found not only in the relatively larger area of the strip cathode, but also in the relatively larger penetration of fields from the anode to the cathode surface. Also, for a given area of transverse section of the beam and for a given beam current and speed, the disturbance of potential along the beam path due to space charge will be less with a beam of rectangular section than with one of circular section.

The merits of strip systems for the production of large beam currents with relatively small space-charge disturbances have been recognized by Broadway and Bull (1940) who recommended the use of strip beams in high-frequency velocity modulation tubes. In these tubes it is desirable that electrodes for extracting power at high frequency should closely embrace the beam currents, which must themselves be large. Usually a number of such electrodes is necessary, which have to be spaced some distance from each other. All the electrodes have relatively small apertures through which the electron beam has to pass in succession. Since, due to the mutual repulsion of the electrons, it is impossible to produce parallel beams of high current, it is the usual practice to direct an initially convergent beam towards a virtual focus which, however, due to space-charge effects, is never reached by the electrons. The actual beam forms a minimum cross-section or a waist, after which it spreads again. The beam waist has to be located in a certain position with respect to the electrodes of the tube if the passage of a maximum current is desired. The spreading of electron beams due to the influence of space charge and the compensation of it by proper field arrangement in the emission system has been treated by Pierce (1939 and 1940). We shall discuss in the following paragraphs a number of experiments which have been set up with a view to studying the current distribution in strip cathode systems so that the influence of electron optics and of space charge and their mutual interaction can be recognized.

§ 2. SPACE CHARGE AND ELECTRON OPTICS IN SIMPLE STRIP SYSTEMS

The mutual influences of space charge and electron optics upon the electron emission can be studied best in very simple strip systems, in which the two effects can be studied independently and separately as far as possible. One of the simplest available systems of this kind is shown in the side sketch of figure 1. It consists of a plane cathode (Ca) and of two equal slots (Gr) and (An). The sketch represents a cross-section in the (yz)-plane in which the beam geometry can be discussed as a two-dimensional problem. All electrodes may be imagined to be extended in the x-direction to such an extent that the ends of the electrodes are sufficiently remote from the (yz)-plane under consideration for the disturbances caused by end effects to be neglected. If the slot apertures are sufficiently small in comparison with their mutual distance (a) and also with the (Gr) slot-to-cathode distance (c), the electron-optical effect of each slot upon the beam can be treated separately in a simple manner.

The focal length (f) of the line-focus lens formed by a slot is given by Davisson and Calbicks' well-known formula

$$f=\frac{2V}{E'-E}, \qquad \ldots\ldots(1)$$

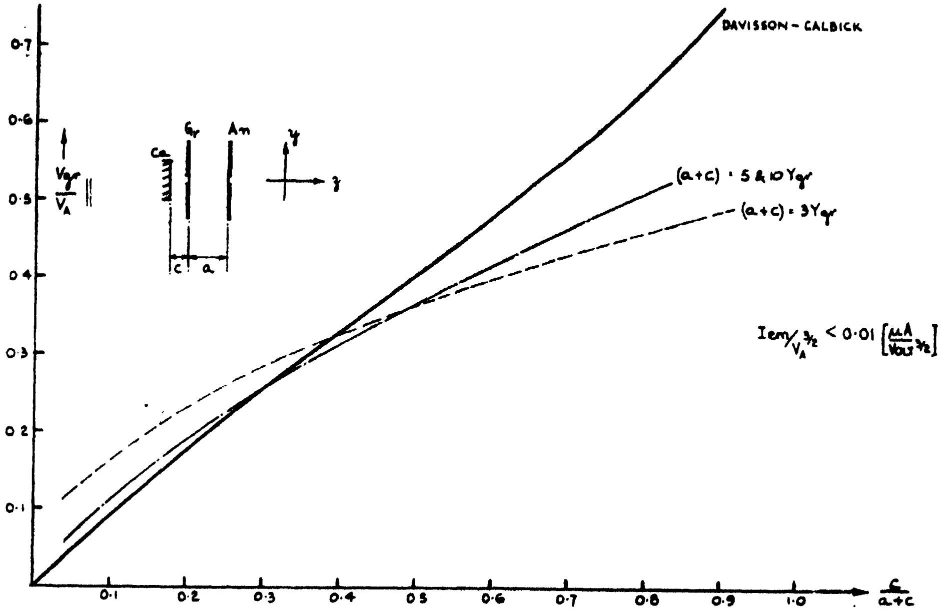


Figure 1. Production of quasi-parallel beams by positive grid bias.

where E and E' are the electric fields (volts/cm.) in the two spaces which are separated by the slot diaphragm, and V is the potential of the diaphragm. C. S. Bull (1945) has investigated recently the application of this formula to aligned grids of thermionic valves, and he has found that the electron paths through the grid apertures were well predicted by the formula. To investigate the application of equation (1) to strip-cathode emission systems, a particular case may be chosen here in order to simplify the experimental tests. According to Bull, the system should emit approximately parallel rays when the gird voltage is positive and reaches a certain fraction of the anode voltage. Calling this fraction

$$n=\frac{V_{gr}}{V_A},$$

formula (1) yields for the grid slot

$$f_1=\frac{2nV_A}{\frac{V_A(1-n)}{a}-\frac{nV_A}{c}}, \qquad \ldots\ldots(2)$$

a and *c* being grid-to-anode and grid-to-cathode spacings respectively.

For the anode slot

$$f_2=\frac{2V_A}{-V_A\left(\frac{1-n}{a}\right)}. \qquad \ldots\ldots(3)$$

The condition for parallel rays is

$$f_1=f_2+a. \qquad \ldots\ldots(4)$$

From equations (2) to (4) it follows that

$$n=\frac{\left(6+3\frac{a}{c}\right)\pm\sqrt{9\left(\frac{a}{c}\right)^2+24\frac{a}{c}}}{6+2\frac{a}{c}}. \qquad \ldots\ldots(5)$$

The voltage ratio *n* as a function of the ratio of spacings $c/(a+c)$ which produces parallel rays according to equation (5) is plotted as a solid curve in figure 1. Only the negative root is considered, since, for practical purposes, we are interested only in values of *n* less than 1. As a purely electron-optical result derived from equation (5), the first slot should always have positive potential for producing parallel rays. On the other hand, experimental evidence shows that under ordinary conditions, i.e. under space charge, a two-slot system always produces divergent beams, the angle of divergence being a minimum when the grid is at zero or at negative bias. It follows, therefore, that under normal conditions of space charge, the mutual repulsion of the electrons is responsible for the divergence of the beam.

The magnitude of space-charge effects in an electron beam of the current (I) with the energy (V) is controlled by the "space-charge factor" ($I/V^{3/2}$). We derive this from Child's law, according to which, in emission systems of any geometry, the electron current grows about proportionally with the 3/2 power of the anode voltage as long as full space-charge conditions are maintained. If we want to get rid of space-charge effects, we have to reduce the emission current (I_{em}) or to increase the anode voltage (V_A) of our system in such a way that $I_{em}/V_A{}^{3/2}$ is reduced. A reduction of $I_{em}/V_A{}^{3/2}$ can be obtained by underheating the filament of the cathode, i.e. by decreasing the cathode temperature.

If in this way the space-charge factor of the simple two-slot emission system was reduced to less than 1/100 of its full value, it was found experimentally that the ordinary electron-optical laws expressed by equations (1) and (5) could be applied. In the experimental test of equation (5) the ratio of grid to anode voltage

$$n=\left.\frac{V_{gr}}{V_a}\right\|$$

for producing parallel rays had to be found as a function of the spacings *a* and *c*. A fairly good estimate of the angle of divergence, or, in particular, of the parallelism of the electron beam, could be obtained with a fluorescent target at a sufficiently large distance from the anode slot of the emission system.

In the actual measurements, the bias at the grid diaphragm was varied until the least beam divergence could be observed. This was taken as an indication of the production of approximately parallel beams. The ratio of this particular grid bias to the anode voltage ($=V_{gr}/V_A\|$) is plotted in figure 1 against the ratio of cathode to grid spacing (c) over cathode-to-anode spacing ($a+c$). The broken curve applies to an ($a+c$) spacing equal to three grid-slot-semi-apertures y_{gr}. The full curve nearest to the dotted one represents the theoretical values of equation (5). Complete agreement with the experimental values could not be expected, since the experimental grid electrode was of finite thickness (1/3 of the width of the grid-slot-aperture); this thickness, however, is neglected in Davisson and Calbick's formula. However, there seems to be no doubt that under the conditions of figure 1, the paths of the electrons are controlled entirely by electron-optical laws.

The measurements refer to the electrode potentials to produce quasi-parallel beams. These beams were not sharply defined but the intensity was strongest in the middle and faded away to the edges. The intensity distribution in the beam is produced by the velocity distribution of the emission and the field distribution at the cathode surface. The velocity distribution is known to be Maxwellian; the field distribution can be calculated or measured in the electrolytic trough. The actual measurement of this distribution and its comparison with the theory would be laborious and of little interest.

Practical emission systems are very rarely used under conditions of extremely low space charges. With increasing space-charge factors, however, the minimum obtainable beam spread increases rapidly. The grid bias, at which the minimum spread is obtained, changes with increasing space charge from positive to zero and to negative values. The intensity of the beam is strongest in the middle, and towards larger angles it appears to fade away so gradually that the edges are not clearly defined.

The current distribution of a beam originates at the cathode, and the emission distribution there may reach a certain equilibrium in its interaction with a particular space-charge distribution. The current distribution of an intense beam is, at a later stage, still decisively influenced by the mutual repulsion of the electrons on their paths. The most essential results on the beam spread can certainly be derived from this mutual repulsion of the electrons. We shall proceed, therefore, to give a short outline of the simple space-charge theory of the beam spread, and we will see how far it can be applied to an interpretation of the experimental results obtained with simple strip systems.

§ 3. APPLICATION OF SIMPLE SPACE-CHARGE LAWS TO STRIP SYSTEMS

A simple theory of the spread of ribbon-shaped beams has been developed by A. Bouwers (1935) and by J. Thompson and L. B. Headrick (1940). This theory can be applied to all parts of the beam in which the forward velocity of the electrons is constant, and in which the direction of the electrons is strictly homocentric in the yz-plane, in which the problem can be treated as a two-dimensional one. There, each electron path describes a parabola, the parameter of which depends upon the initial convergence and upon the space-charge factor ($I/V^{3/2}$) of a beam, the boundary of which is defined by the parabola considered. According to Bouwers, we can distinguish the three main types of paths which are shown in figures 2, 3 and 4 of the present paper.

In all three figures, the beam (El) enters at the co-ordinate (z_A) through a slot of the semi-aperture (y_A). At the slot the beam is directed at an angle θ with the z-axis towards a virtual focus at z_F, which, however, is never reached because of the mutual repulsion. In figure 2 the beam reaches a waist, i.e. a minimum semi-aperture y_W at z_W, while the beam in figure 4 forms a real cross-over with zero aperture at z_X. Figure 3 shows the border case or transition between the types of figure 2 and figure 4. The point of minimum cross-section ($y_W = 0$) at ($z_W = z_X$) can be considered to be the transition between a waist and a cross-over. Its distance from the aperture is here a maximum. The equation of the electron path is

$$y = y_A - \theta z + \frac{1}{4k} I/V^{3/2} z^2. \qquad \ldots\ldots(6)$$

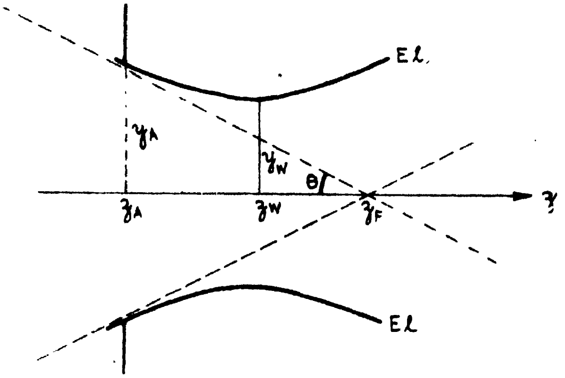


Figure 2. Electron path and space charge.

The aperture being taken as the origin, i.e. $z_A = 0$, from equation (6) follows

$$\left. \begin{aligned} y_W &= y_A - \frac{\theta^2 k}{I/V^{3/2}}, \\ z_W &= \frac{2k\theta}{I/V^{3/2}}, \end{aligned} \right\} \qquad \ldots\ldots(7)$$

and with $y_X = 0$

$$z_X = \frac{\theta \pm \sqrt{\theta^2 - \dfrac{y_A I}{k V^{3/2}}}}{(I/V^{3/2})/2k}, \qquad \ldots\ldots(8)$$

where k is a constant. The numerical value of this constant is 10·4 if the current (I) is measured in microamp. per cm. length in the x-direction, i.e. for a width of the strip-beam $2x_A = 1$ cm. Putting $y_W = 0$ in equation (7) gives for the critical case of figure 3 the following critical space-charge factor:

$$I/V^{3/2} = \frac{k\theta^2}{y_A}. \qquad \ldots\ldots(9)$$

Moreover, since $\theta = y_A/z_F$, it follows from equation (7) for any beam convergence

(θ) and any semi-aperture (y_A) satisfying the critical condition equation (9), that the waist distance from the aperture must be just twice as large as the distance of the virtual focus, i.e. $z_W = 2z_F$.

Figures 2 to 4 are all drawn to scale in the y co-ordinate, but scaled down by a factor 10 in the z co-ordinate in order to make the essential features more visible. The space-charge factor in all these curves is taken as $I/V^{3/2} = 0{\cdot}1$ (microamp./volts$^{3/2}$), but the angle (θ) is chosen as 0·05, 0·1 and 0·12 radians respectively in the three figures. The values of y_W, z_W and z_X, plotted in the figures, are calculated according to equations (7) and (8), the semi-aperture (y_A) being taken as unit length. While figures 2 and 3 need no further explanation, it may be pointed out that the broken curve in figure 4 represents the parabola given by equation (6); this is continued in the drawing after the z-axis is reached at z_X. The real electron path, however, is shown as the solid line (El); it is the reversed first branch of the parabola between (y_A, z_A) and (0, z_X), plotted in ($-y$) direction. It may be emphasized here that the rays in strip-beams of sufficiently large angle of convergence (θ) and sufficiently small space-charge factor ($I/V^{3/2}$), as shown in figure 4, really cross over. In distinction to this, a real cross-over cannot be obtained in beams of circular cross-section, unless the beam currents are so much decreased that discontinuities of space charge due to its composition of single electrons start to play a part.

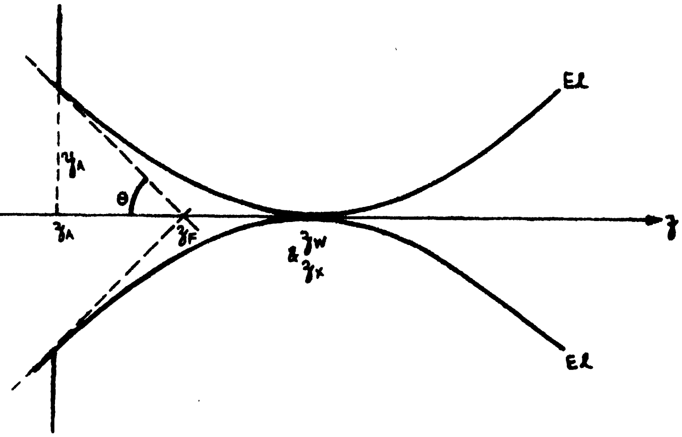


Figure 3.

In the attempt to compare Bouwers' simple theory and the experimental results obtained with simple strip-cathode systems, we are obviously not interested in details of the angular distribution of the emission caused by the field distribution over the cathode surface. On the other hand, if some function of the space-charge factor could be found to represent the beam spread, at least a qualitative comparison would be possible between theoretical and experimental results. In a somewhat arbitrary way we have chosen two characteristic measures for the beam spread which apply to parts of the beam sufficiently far away from the waist. There, the beam has spread sufficiently, so that space charge has hardly any further effect upon the paths of the electrons, which approach straight lines. We have measured at this point either the semi-vertical angle (θ_h), which includes half the beam current, or we have measured the fraction of the total emission I_{Fa}/I_{em}, which is projected into a semi-vertical angle of 0·1 radians.

For both types of measurement an adjustable slot (Sl) in front of a Faraday cage, as shown in figure 5, was used. There, (Ca), (Gr) and (An) represent the strip-cathode, the slotted grid electrode and the anode respectively.

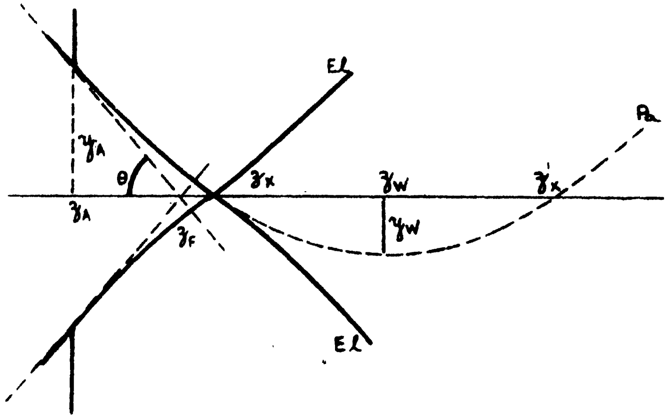


Figure 4.

The slot (Sl) was adjustable in the vacuum; its design was developed from an optical model described by Sears (1933) and by Strong (1941). A schematic diagram is given in figure 6. The brass frame (Fr) was mounted on the four-rod assembly (Rd) on to which the whole gun was fixed. On this frame were pivoted two pairs of parallelogram arms (Pa) holding the jaws (Jw) on to which were screwed the two thin slot-blades (Bl). The jaws were pushed forward against the spring-blades (Sp) by a sliding part (Bp) which was guided in the two blocks (Ga) and (Gb). In order to vary the slot width, the sliding part was screwed forward by the internally threaded helical gear-wheel (Ha) which was rotated and driven by another helical gear-wheel (Hb), arranged at right angles to it. The wheel (Hb) was fixed to an axle which was connected to a ground joint so that it could be turned from outside the tube. The resulting movement of the jaws covered changes of semi-aperture up to $y = \pm 4$ mm. The zero-slot width was adjustable by the position of the blades which were screwed on the jaws.

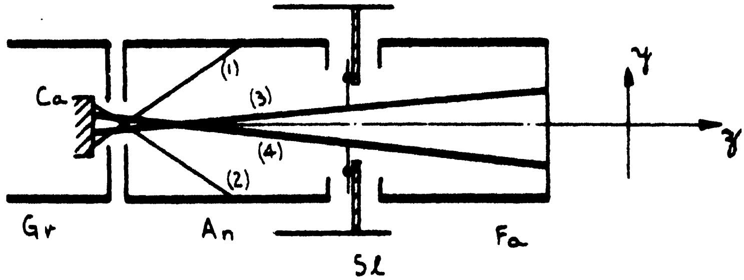


Figure 5. Measurement of angular current distribution.

Experimental curves of the current efficiency, i.e. (I_{Fa}), as a function of (θ) for small θ did not deviate much from a straight line (up to $\theta = 0{\cdot}1$). Only for large θ these curves gradually bend round at some point to converge towards the 100 % efficiency, when all available electrons are included inside the considered wedge.

Figure 7 shows the change of beam spread caused in a given system by the change of the space-charge factor only, the grid electrode being constantly kept at cathode potential. The change of $I_{em}/V^{3/2}$ is plotted as abscissa, this change being produced by a variation of the cathode temperature. The beam spread is characterized by the angular semi-aperture (θ_h) into which half of the total emission is projected. (θ_h) is plotted as ordinate. For small space-charge factors, the electrons follow, as we have seen, purely electron optical laws. The equipotentials near the grid electrode are strongly curved, thus the electron paths cross over steeply and the resulting beam is very divergent. With increasing space charge, the repulsion of the electrons at the cathode side of the cross-over tends to reduce the initial convergence of the beam and thus reduces its final divergence. Thus the left-hand part of the curves of figure 7 is explained by the transition of the beam from the stage characterized by figure 4 to the stage shown by figure 3. The curves of figure 7 pass through a minimum and rise again when, by increased space-charge effects, the initial convergence is further reduced with a simultaneous increase of the divergence after the beam has passed the waist. These conditions correspond to the stage represented by figure 2.

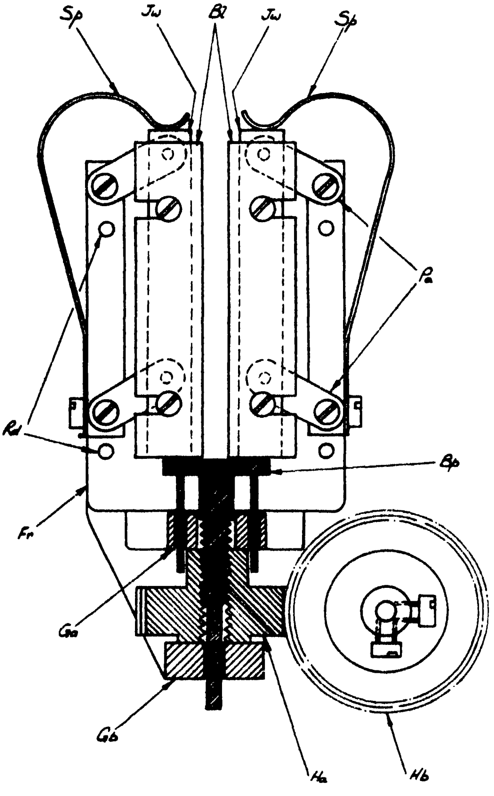


Figure 6. Parallelogram slot.

Bouwers' simple theory, as illustrated by figures 2 to 4, is adequate for a rough qualitative explanation of the essential facts. To understand details about beam spread and angular distribution, the following considerations have to be given:—

1. In every cathode system the space-charge factor ($I/V^{3/2}$) of the beam decreases gradually on the way from the cathode, where (V) is very small, to the anode, where $V = V_A$. Bouwers' theory, however, applies to beams with homogeneous space-charge factors.

2. It has already been pointed out in § 2 that the current distribution of the beam originates at the cathode surface, which is exposed to a certain distribution of field strength. For intense beams, however, this field distribution is produced not only by the electron-optical potential distribution but also by the space-charge distribution in front of the cathode. Eventually, the emission distribution reaches a certain equilibrium in its mutual interaction with the space-charge distribution.

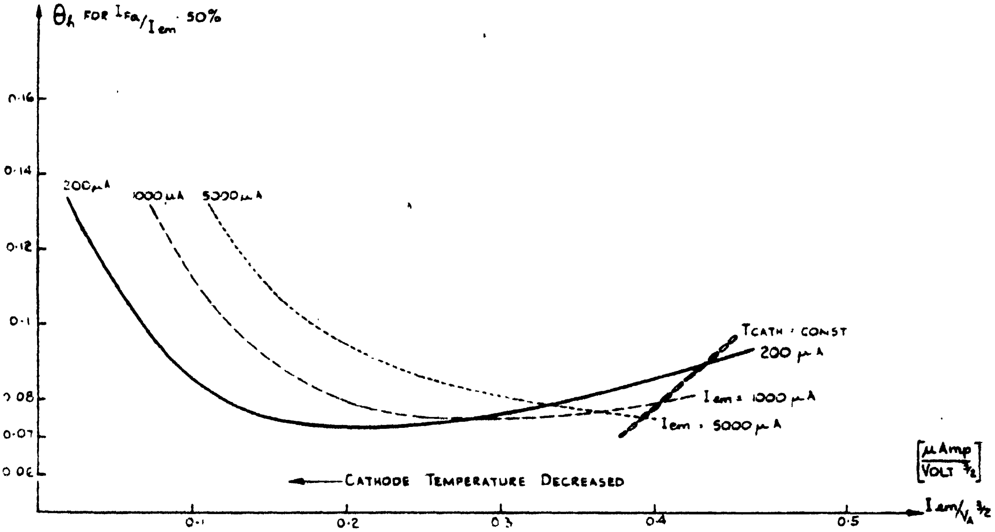


Figure 7. Beam spread and space charge open anode system.

3. The initial current distribution of the electrons at the cathode is modified by an initial thermal-velocity distribution and by the influence of space charge upon this distribution (*cf.* Klemperer, 1947). The influences arising from the velocity distribution, however, are of secondary importance for the current distribution and they will not be studied in the present paper.

4. The potential distribution set up in the beam by the space charge produces a decrease of electron velocities in the paraxial parts of the beam. This reacts again upon the potential distribution until equilibrium is reached. We shall deal with this effect in § 5 of this paper.

Changes in angular current distribution produced by changes in field distribution at the cathode surface as mentioned under (2) above are very noticeable in the transition region between temperature saturation and space-charge condition as in figure 7. There, the current distribution of the beam not only changes with a change in ($I_{em}/V_A^{3/2}$), but even if this latter space-charge factor is kept constant, it changes with the anode voltage (V_A). There is on the cathode surface a central zone with temperature-limited emission. The width of this zone apparently depends upon the absolute value of the field. As a consequence, the three curves of figure 7 which were measured at 0·2, 1 and 5 milliamp. differ appreciably. The change of field distribution at the cathode surface is also responsible for the fact that at a constant cathode temperature, and even under conditions of full space charge, the factor ($I_{em}/V_A^{3/2}$) decreases appreciably with increasing anode voltage or emission current. A short curve, which in figure 7 is indicated by a chain of small circles, connects the points of equal cathode temperature under full space-charge conditions in the curve of beam spread against space-charge factor.

The curves of figure 7 are instructive in showing the powerful influence of the gradually increased space charge. In practical applications, however, emission systems are always used under full space-charge conditions. There, both the full space-charge factor at the anode, which is usually called the *perveance* of the system, and the divergence of the emitted beam can largely be controlled by the spacing between cathode and grid electrode (c), by the spacing between grid electrode and anode (a), and by the widths of the semi-apertures (y_{gr}) of the grid slot and (y_A) of the anode slot.

Experimental values for the fraction of the total emission under full space-charge conditions that is projected into an angle of $\theta = \pm 0{\cdot}1$ radians are shown by curves in figure 8 for a few characteristic representative systems. These systems were chosen from the great number of possible combinations of plane, slotted diaphragms and of a plane cathode. This current fraction characterizes the efficiency of the emission system. It is, for instance, high (0·8) for a system with an open, wide anode (no diaphragm in the anode) and low (0·24) for a symmetrical system with equal grid and anode slot (semi-apertures $y_A = y_{gr}$) having an anode-to-grid-spacing equal to one grid slot semi-aperture ($a = y_{gr}$). The efficiency of both these systems is not markedly dependent on the cathode-to-grid spacing (c).

It would be expected that for decreased spacing (c), the electron-optical effect of the smaller curvature of the equipotentials near the grid electrode would produce a smaller initial convergence. This effect, however, seems to be balanced to some extent by an increase in beam spread due to an increasing space charge, the space-charge factor being larger the smaller the spacing (c). If, now, with further decrease of (c), the space-charge factor is further increased, beam conditions will eventually reach the critical stage represented by equation (9). We can then expect a minimum in beam spread, i.e. a maximum in efficiency. The critical spacings of the electrodes, at which this maximum occurs, have been measured and the particular conditions have been discussed sufficiently in a paper on emission systems with circular symmetry by Klemperer and Mayo (1947), so that we need not enlarge further on this subject. We shall only point out here some characteristic differences between the circular and the strip systems.

It has been pointed out elsewhere (*cf.* Klemperer, 1939, p. 96) that two-dimensional (line focus) lenses always have shorter focal lengths than three-dimensional circular-symmetry lenses of the corresponding electrode arrangement. Thus the critical conditions which for symmetrical circular two-diaphragm systems were found to occur at a critical grid-to-cathode spacing of the order of a grid semi-aperture ($c \approx y_{gr}$) would be expected for the corresponding slot system at much smaller spacings. This is borne out by the two dotted curves of figure 8, which do not show maxima within the plotted range. Apparently, for technical reasons, the spacings (c) from the top of the grid to the cathode could not be made small enough to reach the critical conditions.

On the other hand, an unsymmetrical two-slot system, with an anode slot half as wide as the grid slot ($y_A = y_{gr}/2$), would have electron-optically a greater focal length and, in addition, would have greater perveance ($I_{em}/V_A{}^{3/2}$) than the corresponding symmetrical two-slot system. Now, grid-to-cathode spacings leading to the critical conditions equation (9) can be easily realized. The broken curves in figure 8 belong to such an unsymmetrical two-slot system and they show distinctly the expected maxima. The three curves shown were all taken under full space-charge conditions, but at the three different emission currents of 1, 5 and 20 milliamp. The wide difference between these three curves points to a relatively large difference in field distribution over the cathode surface, which seems to have a particularly great effect upon the angular current distribution under the critical conditions near the maximum efficiency. In this respect, slot systems behave analogously to the circular systems; however, in contrast to the circular systems, even near the critical stage, the slot systems are far less subject to the influence of residual gas.

Concerning the systems with open box anode the critical conditions of equation (9) may be obtained by reducing appreciably the dimensions of the anode. This can be seen by comparison of the two full curves of figure 8; no further comments, however, need be given about this case.

As a point of practical interest, it may be mentioned that the perveance of the different types of slot systems, even for a given length of slot, may be of a very different order. The perveance of course changes very rapidly with the spacings a and c, but even if we fix both these spacings at $1y_{gr}$, we obtain the following rather different values: $I_{em}/V_A{}^{3/2} = 0{\cdot}1$ μamp./(volt)$^{3/2}$ per cm. slot length for the slot system with wide, open anode, but 5 μamp./(volt)$^{3/2}$ per cm. slot length for the unsymmetrical two-slot system. In comparison, the circular symmetrical two-diaphragm systems and open anode systems reach only $I_{em}/V^{3/2} = 0{\cdot}2$ and $0{\cdot}006$ μamp./volt respectively.

§4 EMISSION SYSTEMS OF PIERCE'S TYPE

The discussions of §§ 2 and 3 apply to the simplest type of strip-cathode emission systems consisting of plane electrodes only. It was shown how, in these systems, space-charge effects completely upset the particular orbits of the electron rays that follow from purely electron-optical laws, and that are realized only at vanishingly small space-charge factors. The field distribution at the cathode surface further complicates the angular current distribution in the beams emitted from the above simple systems.

In order to have theoretically simple conditions which would allow the highest possible beam concentrations, Pierce (1939 and 1940) proposed a new kind of emission system in which the electric field is constant over the whole cathode surface and in which space-charge effects would not tend to alter the electron-optical paths of the electrons. In Pierce's strip system, cathode and grid electrode are parts of two coaxial cylinders. It is known that between two complete coaxial cylinders a strictly radial flow of electrons is obtained whatever conditions of space-charge are given. If a sector is taken out of the complete cylinder, the radial flow of the electrons can be preserved if the potential distribution along the beam is kept unchanged everywhere, especially at the boundary face, which now occurs between electron beam and empty space. Pierce suggested producing at this boundary—by means of suitably shaped electrodes—a field which agrees with the space-charge conditions, i.e. which shows along this boundary a potential distribution proportional to the 4/3 power of the distance from the cathode.

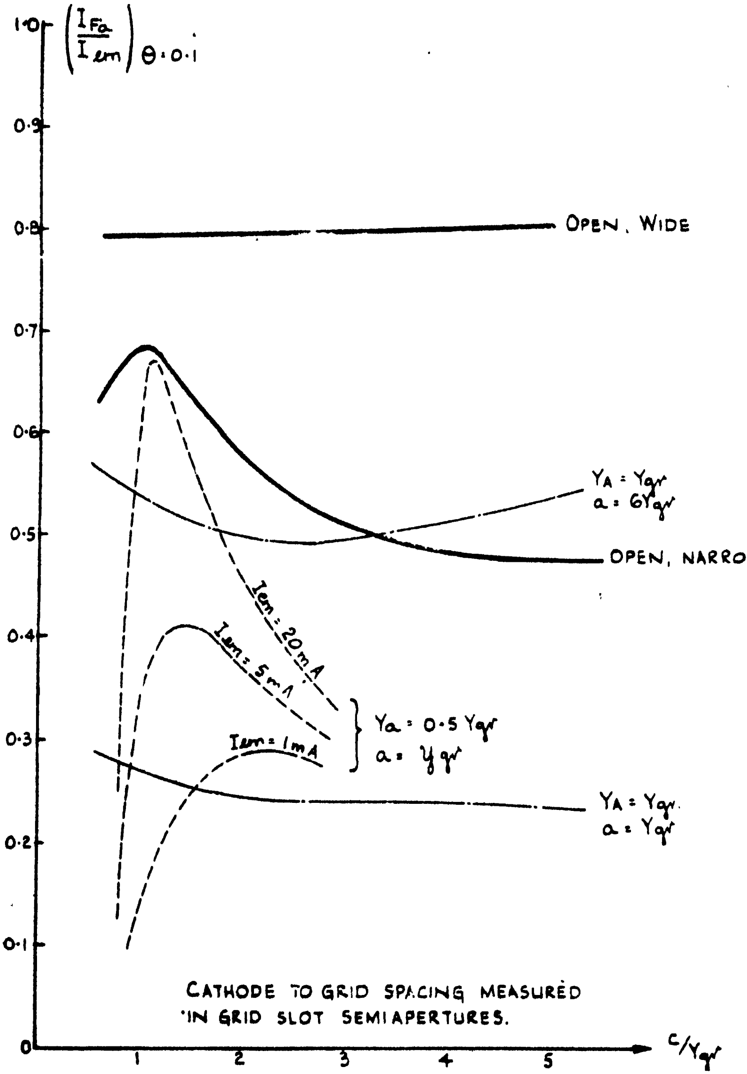


Figure 8. Efficiency = fraction of total emission projected into an angle $\theta = \pm 0 \cdot 1$. Full space charge.

According to Pierce, fields of this kind can be practically obtained with the help of the field-plotting trough. As far as the boundary is concerned, the electron beam can be replaced there by a piece of insulator, since the normal field component in the electrolyte at the boundary of the insulator will vanish just as the normal field component in the vacuum vanishes at the edge of the electron beam. Figures 9 and 10 show two different cathode systems, both having been shaped according to experiments with large-scale models in the field-plotting trough so as to satisfy Pierce's conditions.

The system shown in figure 9 has a concave cathode (Ca) which is surrounded by the cathode shield (Sc), both sides of which are fixed to the two field-shaping

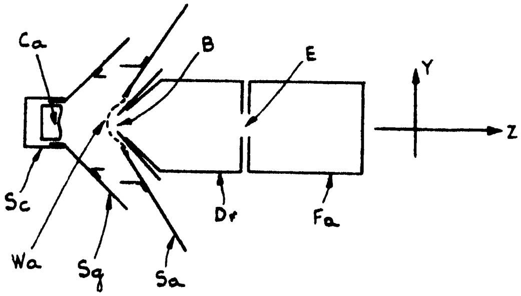


Figure 9. "Pierce gun" and "electron tunnel."

ls (Sg). Opposite to (Ca) there is arranged a wire grid anode (Wa) o
lar cross-section in the yz-plane, which bears the two field-shaping shields (Sa)
rons, all homocentric in the yz-plane, are supposed to pass through the wir
Wa) and to enter the field-free tunnel (Dr) through the slot (B); they leav
through a slot (E) equal in size to (B).

e should expect to get a maximum current through a tunnel formed by tw
slots if the parabolic electron orbits were arranged symmetrically to it, i.e

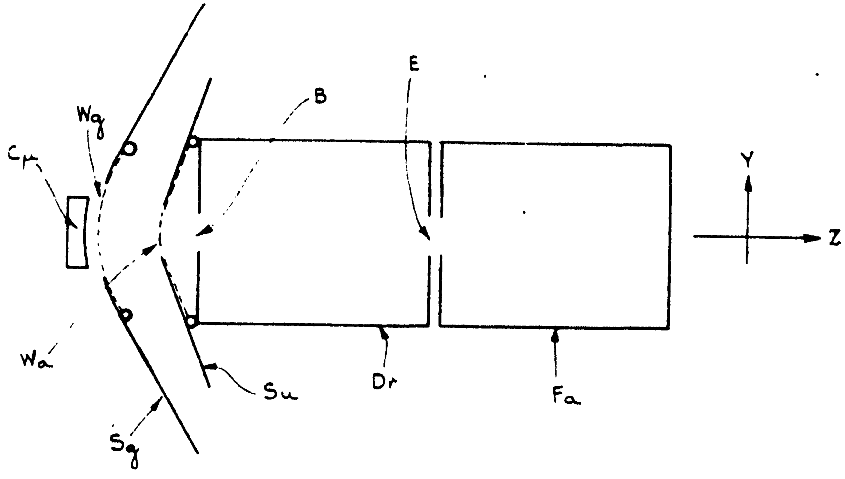


Figure 10. "Pierce gun" and "electron tunnel."

if the beam forms its waist exactly in the middle of the tunnel. Moreover, space-charge factor and angle of convergence (θ) at the aperture should be chosen so as to produce the critical case represented by equation (9) or figure 3. Now θ is fixed by the geometry of the tunnel, i.e. by the width of the slots ($2y_A$) and their mutual distance ($2l$). Since, as we explained above for the critical case, the distance of the virtual focus from the aperture (z_F) is just one-half of the waist distance ($1/2\,z_W$) from this aperture, we have

$$\theta = \frac{y_A}{z_F} = \frac{2y_A}{z_W} = \frac{2y_A}{l} \qquad \ldots\ldots(10)$$

and with equation (9)

$$I/V^{3/2} = \frac{8ky_Ax_A}{l^2}, \qquad \ldots\ldots(11)$$

where $2x_A$ is the breadth of the slot, which is not written down expressly in the notation of equation (9), since this equation applies to the unit beam breadth.

If we require the total output of the cathode system to pass through the tunnel, the system has to satisfy the requirements given in equations (10) and (11). Bull and Klemperer (1944) drew the further conclusion that in cathode systems such as shown in figures 9 and 10, the curvature of the cathode and the cathode-to-anode distance are both fixed by equations (10) and (11) as soon as the tunnel dimensions are given. This can be seen immediately from Child's equation which—according to B. J. Thompson (1943)—can be written for a cylindrical cathode system, expressing again the total emission I_{em} in microamp. and the anode voltage V_A in volts:

$$\frac{I_{em}}{V_A^{3/2}} = 2{\cdot}33\,\frac{A}{r_A^2\beta^2} \qquad \ldots\ldots(12)$$

where A, the area of the cylindrical anode, is given accurately enough by $(2y_A)$ as long as the arc can be replaced by the chord $(2y_A)$. Moreover, r_A represents the radius of curvature of the anode which is here $r_A = z_F$. β is a function of (r_A/r_c), the ratio of the radii of curvature of anode and cathode. The function β is given in a graphical representation by I. Langmuir and K. I. Compton (1931), where it is approximated by the following series:

$$\beta = \log_e\frac{r_A}{r_c} - 0{\cdot}40\left(\log_e\frac{r_A}{r_c}\right)^2 + 0{\cdot}092\left(\log\frac{r_A}{r_c}\right)^3. \qquad \ldots\ldots(13)$$

On the other hand, β is given by substitution of $I/V^{3/2}$ of equation (11) for $I_{em}/V_A^{3/2}$ into equation (12) since the cathode system is supposed to produce a beam of exactly the space-charge factor which the tunnel requires. Thus

$$\beta^2 = \frac{2{\cdot}33}{z^2_F}\,\frac{(2y_Ax_A)(2z_F)^2}{(4y_Ax_A)k} = 0{\cdot}448$$

which, with equation (13), yields

$$\frac{r_c}{r_A} = 1{\cdot}71 = \frac{z_F + d}{z_F} \quad \text{or} \quad \frac{d}{z_F} = 0{\cdot}71,$$

i.e.

$$d = 0{\cdot}35l, \qquad \ldots\ldots(14)$$

where $d = r_c - r_A$ is the distance between cathode and anode.

It follows that we should expect to pass the total emission of the cathode through a tunnel of relatively small aperture as soon as (1) the anode radius is a quarter of the length of the tunnel, (2) the cathode-to-anode spacing is 0·17 times the length of the tunnel.

The system shown in figure 9 was assembled in exactly these required geometrical proportions. As a further precaution, the strip cathode (60 × 6 mm., radius 10 mm.) was covered up at the ends to cut out the regions of disturbance according to a scheme tried out in earlier experiments. $I_{em}/V_A^{3/2}$ was as expected, and its measured value was about 25. Less than 20 % of the cathode emission, however, was received in the Faraday cage (Fa). Only a few per cent of this emission was caught at the wire grid (Wa), which consisted of 60 parallel 0·1 mm. dia. molybdenum wires welded across the anode slot and at the jaws of the entrance slot (B) of the tunnel. Most of the emission current was apparently caught at the diaphragm containing the exit slot (E) of the drift space. Improved yield (35%) could be obtained by changing the cathode-to-anode distance to an optimum of 6 mm., reducing thus the perveance $I_{em}/V_A^{3/2}$ to 12 microamp./volt$^{3/2}$.

There was some reason to suspect disturbances of the field caused by the unavoidable gap between the grid shield (Sg) and the cathode. These disturbances were avoided by another construction shown in figure 10. All inscriptions at this figure correspond to those in figure 9. However, a grid (Wg) (again consisting of 60 molybedenum wires of 0·1 mm. dia. welded in 1 mm. mutual distance across the slot) was inserted closely in front of the cathode.

The grids were pressed to the correct radii of curvature, with the help of specially made jigs, in order to obtain the necessary accuracy in agreement with the theoretical requirements. The single wires of the grids (Wg) and (Wa) were aligned, and in order to avoid undesirable electron focusing by Wg, this grid was kept at a slightly positive bias, which was found empirically by adjusting it to obtain the best yield in the Faraday cage. The current efficiency of this system (figure 10) was not better than that of the system of figure 9.

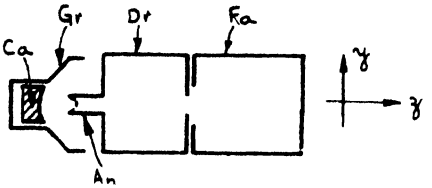


Figure 11. Empirical cathode system for electron tunnel.

When we varied the radii of curvature of the grids Wg and Wa, we noticed, however, that the efficiency could be improved to 60 % of the total emission when the grids were not coaxial, but the radius of curvature of Wg was made smaller and the radius of curvature of Wa larger than their respective distances from the required virtual focus, z_F.

In the end, by a purely empirical development, we arrived at the construction shown in figure 11. There, concave cathode (Ca) and grid shield (Gr) were similar to the corresponding electrodes in figure 9. The entrace of the tunnel (Dr) was formed by the anode nozzle (An). The cathode system had a virtual focus near the end of this anode nozzle. Inside the nozzle, however, there was a strongly diverging field produced by concave equipotentials penetrating into it. Due to the simultaneous action of this diverging field and of the mutual repulsion inside the beam, a waist was probably formed in the middle of the drift space. The optimum anode-to-cathode distance was found empirically by shifting the cathode and grid electrode (Ca and Gr) by a micrometer bellows gear and observing the current in the Faraday cage (Fa). This current could also be adjusted to an optimum by varying a small voltage applied to the grid electrode. This bias voltage was found to equal cathode potential for the above optimum and was found to be negative for shorter distances and positive for larger distances. Apparently the bias controlled the waist distance from the cathode; probably the waist was formed in the middle of the tunnel for the correct combination of cathode position and bias. Moreover, the application of a fine wire grid inside the anode nozzle has been tried. It was found that no disturbances were caused if the grid was strongly concave towards the cathode, as is indicated by the dotted line in figure 11. If, on the other hand, a convex grid was used, such as shown in the guns of figure 9 and figure 10, the current efficiency was greatly reduced.

From experimental results of this kind, it can be concluded that in order to transmit a beam of the greatest possible space-charge factor through a narrow tunnel of given dimensions, this beam should be caused to be "initially overfocused", i.e. it should aim at a virtual cross-over appreciably nearer to the cathode than would be required by equation (10). Then, entering the tunnel, the beam should be spread out by passing through a series of equipotentials which are concave facing the cathode.

In this way particularly compact emission systems can be designed in which the distance between cathode surface and beam waist is relatively short. However, the shorter this distance can be made the less pronounced will probably be the longitudinal potential distribution in the beam and the beam spread produced by it. The potential distribution in the beam will be discussed in detail in the next section of this paper.

The system of figure 11, which represents an optimum design, had a current efficiency of 80%. This implied that a current of the space-charge factor of 7 microamp./(volt)$^{3/2}$ passed through a tunnel of the dimensions $x_A = 30$, $y_A = 1{\cdot}5$, $l = 10$ mm. This, however, represented only 18% of the space-charge factor (38 microamp./volt$^{3/2}$) which was expected from theoretical reasons to pass through this tunnel according to equation (11). The "Pierce" system of figure 9 passed only 10% of the theoretical value through its tunnel. Moreover, the best empirical system that could be made up from a plane strip-cathode and two plane, slotted diaphragms, such as described in § 2 of this paper (*cf.* maximum of broken curve in figure 8) had an efficiency of 45% through a tunnel of similar proportions, but the maximum transmitted $I/V^{3/2}$ amounted only to 17% of the theoretical optimum expected from equation (11).

The current efficiency is to some extent critical with respect to alignment and adjustment, and the chance that a still more efficient type of system might be found by future research cannot be excluded. However, the great number of results obtained with very different types of systems suggests that even under the most favourable conditions a given tunnel could in practice not pass more than, say, a quarter of the space charge which is calculated by equation (11) on the simple theory outlined in §3. This statement refers to slot systems. Analogous experiments with circular systems lead to similar conclusions. However, the upper limit of the ratio of practical to theoretical space-charge factor of the current passed through a circular tunnel was decidedly better and was estimated to be about 1/2 for tunnels in which diameter and length were of the same order of magnitude.

§5 REASONS FOR THE INADEQUACY OF THE SIMPLE THEORY OF BEAM SPREAD

In the attempt to find reasons for the quantitative discrepancies between the expectations of the simple beam spread theory, outlined in §3, and the results given in §4, certain indications have been obtained from electronic ray-tracing experiments. A divergent beam emitted from a cathode system may be intercepted by a pepperpot diaphragm and the pencils may be traced with microscope and sliding fluorescent target (*cf.* Klemperer and Wright, 1939; also Klemperer and Mayo, 1946). If the cross-over of such a beam is extrapolated, it appears that the rays do not emerge from a common centre. Moreover, if the cathode temperature is changed, the pencils generally become increasingly less homocentric with increasing space-charge factor. Space charge appears to introduce positive spherical aberration. This is explained in figure 12. There, two marginal rays (Ma) and two paraxial rays (Pa) emerging through a pepperpot diaphragm (Pp) are extrapolated backwards. Their virtual cross-overs do not coincide, i.e. the bundles are not homocentric. From these observations, the actual waists (Wm) and (Wp) of the marginal and paraxial rays respectively can be located approximately. Apparently, the marginal waist is formed first and the paraxial waist later.

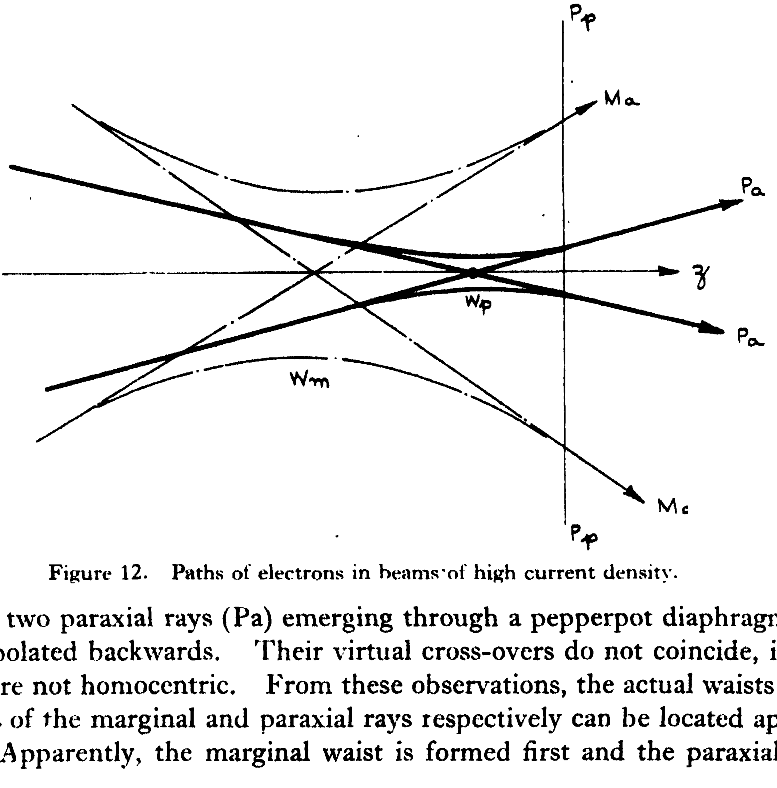


Figure 12. Paths of electrons in beams of high current density.

The diagram in figure 13 represents a potential distribution across the beam. Due to space charge, the negative potential difference $(V_A - V)$ rises beyond the anode voltage (V_A) towards the middle of the beam. This is indicated by the broken line, which could be calculated for a homocentric bundle under the assumption of constant space-charge density over the *xy* cross-section. The potential V_0 at the axis can be derived from Gauss's law and is given by the following equation:

$$V_A - V_0 = \frac{\pi y_A}{\epsilon_0 \sqrt{\frac{2e}{m}}} \frac{I}{V_A^{1/2}} = 4{\cdot}76 \times 10^{-2} y_A V_A \frac{I}{V_A^{3/2}}. \qquad \ldots\ldots(15)$$

where the beam current I is measured again in microamp. per cm. breadth, the potentials V in volts and the beam semi-aperture y_A in cm. ϵ_0 is the dielectric constant. It can be seen that the potential hump in the middle of the homocentric beam grows proportionally with the semi-aperture of the beam and with the beam current, but inversely proportional with the square root of the beam voltage.

Due to this rise in negative voltage, however, the paraxial part of the beam is slowed down, i.e. its space charge is increased with respect to the space charge of the marginal beam. As a consequence, the paraxial potential will be further raised, and so on, until a new equilibrium is obtained which is indicated in figure 13 by the solid curve. It can be seen from the figure that, in the new curve, the paraxial potential gradient has risen more than the marginal one. Moreover, due to the increased paraxial space-charge factor, the position (z_W) of the paraxial waist is shifted according to equation (7). As a consequence of the paraxial voltage change there would be thus expected:

(1) Loss of homocentricity.
(2) Loss of homogeneity of beam current density through the cross-sectional area.
(3) Loss of symmetry of the beam waist in the yz plane.
(4) Dependence of beam geometry upon anode voltage.

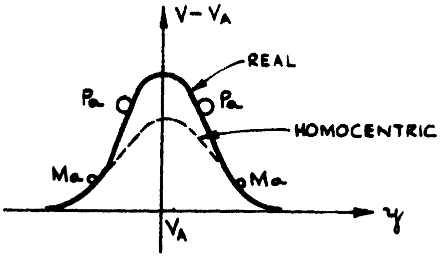


Figure 13. Lateral potential distribution in beam of high current density.

Even if the space-charge factor were kept constant, the paths of the electrons could no longer be considered to be invariant with respect to voltage changes. The paraxial voltage hump would grow according to equation (15), proportionally with the beam voltage, if $I/V_A^{3/2} = \text{const.}$

A theoretical investigation of the potential distribution in an electron beam due to its space charge has been published by Smith and Hartman (1940), and some approximate expressions have been presented for the spreading of circular beams. According to these authors, the velocity distribution of the electrons, due to the space charge, causes the beam to spread more rapidly than can be concluded from the simple theory (*cf.* Watson, 1927). The circular beam will diverge a given amount in about one-half of the distance computed from the simple equations of Watson. Our experimental results, however, neither lead directly to the beam spread curve nor to the potential distribution.

Since a knowledge of the potential distribution in the beam appears to be of basic importance for the understanding of the actual space-charge effects, we have started to investigate it by direct measurements. As a first step, we have obtained qualitative results for a potential distribution across and along a spreading ribbon-shaped beam.

Wehnelt and Bley (1926) first showed that the space charge in front of a cathode can be probed by a fine electron beam. We have adapted their procedure for our purposes and our experimental arrangement is shown in figure 14. There, a very fine strip beam (El) emerging from the anode (An) of an electron gun was used to probe the potential distribution of a space charge emitted from the cathode (Ca). The deflection of the electrons of this probe beam by the space charge was eventually measured on a fluorescent target (Ta). The position of the probe beam was adjusted by the deflection plates (Df). The probe beam while passing the space charge was only about 0·1 mm. wide. A beam velocity of 500 volts was found to be convenient. The probe beam passed through the open sides of an anode-box formed by the anode slot (As) and the anode roof (Ar) of a "space-charge producing" gun. The latter contained a two-slot emission system such as has been described in §§ 2 and 3 of this paper. This system had a plane cathode (Ca), a flat grid electrode (Gr) with a slot aperture and a slotted anode (As). The grid electrode (Gr) was shielded by a wire gauze (Sh) connected to the potential (V_A) of the anode (As) and (Ar) to which was also connected the probe-gun anode (An′), one of the deflector plates (Df) and the fluorescent target (Ta). Behind the target, there are indicated in the figure the two co-ordinate systems (y', z') of the probe gun and (x, z) of the space-charge producing gun.

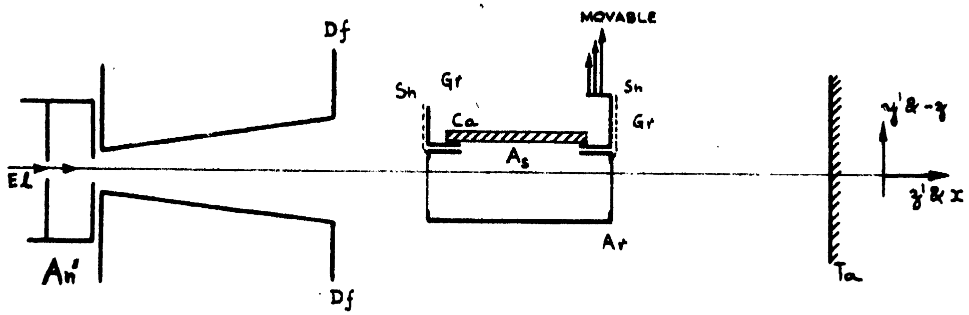


Figure 14. Probing of space charge.

Initially the cathode (Ca) was not heated, so that no space charge was produced in the anode box. As long as grid (Gr) and cathode (Ca) were kept at the potential (V_A), the cross-section of the probe beam could be seen on the target as a straight line, as shown in figure 15 (*a*). When the cathode (Ca) and grid (Gr) were charged up to −20 volts with respect to the anode voltage (V_A), the picture shown in figure 15 (*b*) could be seen on the fluorescent target. The bulge in the middle of the line was due to the bulging of the equipotentials in front of the anode slot (As), since the field between cathode and anode was penetrating through this slot. When, further, the cathode was heated by a proper filament current, its emission current was observed to produce the picture shown in figure 15 (*c*) on the target. The bulge had widened, and moved in the direction of the cathode, while the ends of the line kept approximately their original position. The movement of the bulge at the target was measured by a microscope within 0·1 mm. accuracy and found to be 3 mm. The movement in the anode box could be estimated to be less than 1 mm

The shift of the probe beam in the middle of the bulge is a significant indication for the field set up by the space charge along the z-axis of the anode box. The observed movement of it towards As showed that there must have been set up a potential minimum in the anode box, and that the probe beam was passing at the side of the minimum which gradually sloped towards the cathode.

Now, the whole space-charge producing gun was mounted on a slide, and by means of an external ground joint it could be screwed up and down in the z-direction as is indicated in figure 14 by arrows with the inscription " movable ". Thus, the potential distribution could be explored throughout the anode box by observing the movement of the probe beam on the target.

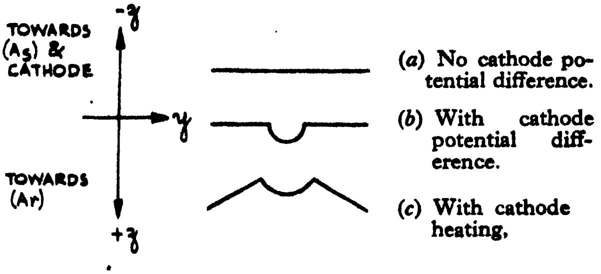

TOWARDS (As) & CATHODE
-z
y
TOWARDS (Ar)
+z
(a) No cathode potential difference.
(b) With cathode potential difference.
(c) With cathode heating,

Figure 15. Picture on fluorescent target.

An experimental result for the longitudinal distribution is shown in figure 16. There, plotted as abscissa, is the position of the probe beam on the z-axis of the anode box; anode slot (As) and anode roof (Ar) position are marked in the graph. The ordinate represents the space potential. The solid curve was taken without space charge, the broken curve with space charge. The close proximity of the potential minimum to the anode roof (Ar) seemed surprising, but can probably be explained by taking into account some secondary emission or electron reflection at Ar.

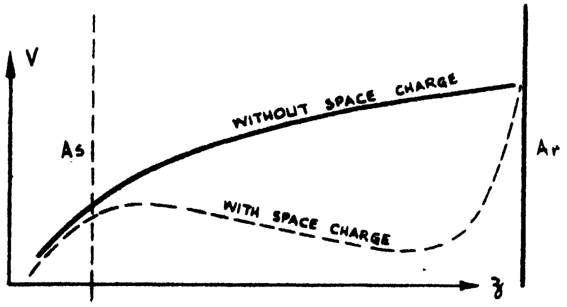


Figure 16. Longitudinal potential distribution in anode space between As and Ar, as measured by a probe beam.

Lateral potential distributions along the various y co-ordinates of the anode box could be estimated from the changes in shape of the line observed at the target, as shown in figure 15. The results lead to curves of the kind that have been shown already in figure 13. The potential minima obtained are deep enough to produce a strong inhomogeneity of electron velocities, The results of this section are only qualitative, but they clearly show two important phenomena, the building up of deep potential minima across and along an electron beam. Further experiments, however, will be needed to explain quantitatively the discrepancy between the experimentally observed current distribution and beam spread predicted by the simple theory.

ACKNOWLEDGMENTS

The work described in this paper was carried out on behalf of the Director of Scientific Research, Admiralty, and the author wishes to thank the Board of Admiralty for permission to publish. The author also wishes to acknowledge his indebtedness to Mr. I. Shoenberg, Director of Research, to Mr. G. E. Condliffe and Mr. L. F. Broadway, and to his colleagues in the Laboratories of Electric and Musical Industries, Ltd., Hayes, Middlesex, where the work was carried out.

REFERENCES

BOUWERS, A., 1935. *Physica*, **2**, 145.
BROADWAY, L. F. and BULL, C. S., 1940. British Patent No. 574,512.
BULL, C. S., 1945. *J. Instn. Elect. Engrs.*, **92**, 86.
BULL, C. S. and O. KLEMPERER, 1944. (Unpublished.)
KLEMPERER, O., 1939. *Electron Optics* (Cambridge : The University Press).
KLEMPERER, O., 1947. "Influence of space charge on thermionic emission velocities." *Proc. Roy. Soc.*, A. (In course of publication.)
KLEMPERER, O. and MAYO, B. J., 1947. "Electron optics and space charge in simple emission systems with circular symmetry." *J. Instn. Elect. Engrs.* (In course o publication.)
KLEMPERER, O. and WRIGHT, W. D., 1939. *Proc. Phys. Soc.*, **51**, 296.
LANGMUIR, I. and COMPTON, K. T., 1931. *Rev. Mod. Phys.*, **3**, 248.
PIERCE, J. R., 1939. British Patent No. 545,835.
PIERCE, J. R., 1940. *J. Appl. Phys.*, **11**, 548.
SEARS, J. B., 1933. *J. Sci. Instrum.*, **10**, 376.
SMITH, L. P. and HARTMAN, P. L., 1940. *J. Appl. Phys.*, **11**, 220.
STRONG, J., 1941. *Rev. Sci. Instrum.*, **12**, 213.
THOMPSON, B. J., 1943. *Proc. Inst. Radio Engrs.*, **31**, 485.
THOMPSON, B. J. and HEADRICK, L. B., 1940. *Proc. Inst. Radio Engrs.*, **28**, 319.
WATSON, E. E., 1927. *Phil. Mag.*, **3**, 849.
WEHNELT, A. and BLEY, H., 1926. *Z. Phys.*, **35**, 338.

A SERIES FOR THE STATIONARY VALUE OF A FUNCTION

BY T. SMITH, F.R.S.,
National Physical Laboratory, Teddington

MS. received 9 *December* 1946

ABSTRACT. A formula is given for the stationary value of a function of any number of variables in terms of the values of the function and its differential coefficients at any point in the neighbourhood of the stationary position.

THE following theorem is to be proved:—

Let f be a function of any number of independent variables x_1, x_2, x_3, ... Denote differentiation of any function with respect to these variables by the addition of the corresponding suffix, so that f_1, f_2, f_3, ... are the first differential coefficients and $f_{11}, f_{12}, f_{13}, \ldots f_{22}, f_{23}, \ldots$ the second differential coefficients of f. It is assumed that the Hessian of f does not vanish, so that the symmetrical matrix

$$\begin{bmatrix} f_{11} & f_{12} & f_{13} & \cdots \\ f_{21} & f_{22} & f_{23} & \cdots \\ f_{31} & f_{32} & f_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

has an inverse **H**. Let the operator **D** be defined by

$$-\mathbf{D}\phi = (f_1 f_2 f_3 \ldots)\mathbf{H} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

(where $\phi_1, \phi_2, \ldots$ are the differential coefficients of the operand ϕ with respect to $x_1, x_2, \ldots$) with the special convention that **D** does not operate on the first differential coefficients of f. Then if the series

$$f + \sum_1^\infty \frac{1}{(n+1)!} \mathbf{D}^n f = F$$

is convergent, the first differential coefficients of F with respect to all the variables vanish, and F represents a stationary value of f.

It will be noted that the values of the variables at the stationary position are not required.

Since every term of F but the first is of the second or a higher order in the first differential coefficients of f, F represents a stationary value of f when these coefficients are zero. When they are not zero the changes in the variables required to approach nearer to the stationary position are, to a first approximation, proportional to these coefficients, so that an expansion in powers of first differential coefficients is appropriate. Normally there is no difficulty in choosing values of the variables which make the first differential coefficients (but not the Hessian) small; convergence is then rapid. The accuracy of the formula may be established by showing that the contribution of any term to a first differential coefficient of F can be written as the sum of two terms, and that the contributions of successive terms cancel one another.

Using literal suffixes to denote differentiation with respect to typical variables, the formula may be written

$$F = f - \frac{1}{2!} f_a f_b f^{ab} + \frac{1}{3!} f_a f_b f_c f^{cp} f^{ab}{}_p - \frac{1}{4!} f_a f_b f_c f_d f^{dq} (f^{cp} f^{ab}{}_p)_q + \cdots,$$

where the convention is adopted that the appearance of a letter in a product both as a subscript and a superscript implies a summation for all variables. Superscribed letters identify elements of the matrix

$$\mathbf{H} = \begin{bmatrix} f^{11} & f^{12} & f^{13} & \ldots \\ f^{21} & f^{22} & f^{23} & \ldots \\ f^{31} & f^{32} & f^{33} & \ldots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

which is necessarily symmetrical as it is the inverse of the Hessian matrix. From the definition of **H**

$$\left.\begin{aligned} f_{ab}f^{bc} &= 0 \quad (a \neq c), \\ &= 1 \quad (a = c). \end{aligned}\right\} \qquad \ldots\ldots(1)$$

Differentiating with respect to a variable g,

$$f_{abg}f^{bc} + f_{ab}f^{bc}{}_{g} = 0.$$

Since subscripts denote differentiations, the first term is unaltered if a and g are interchanged, and therefore the same change can be made in the second term, or

$$f_{bg}f^{bc}{}_{a} = f_{ak}f^{kc}{}_{g},$$

where for convenience different letters are used to denote the dummy variables. Multiply both sides of this equation by $f^{am}f^{n}{}_{g}$. Then

$$(f_{bg}f^{gn})f^{am}f^{bc}{}_{a} = (f_{ak}f^{am})f^{gn}f^{kc}{}_{g}.$$

Since the number of variables is finite we may sum the terms in whatever order we like. By (1) the bracketed terms are zero except when on the left $b=n$ and on the right $k=m$. With these values both bracketed sums are equal to unity, and therefore

$$f^{am}f^{cn}{}_{a} = f^{gn}f^{cm}{}_{g} = f^{an}f^{cm}{}_{a}, \qquad \ldots\ldots(2)$$

or the interchange of the indices m and n does not alter the value of the product.

This result may be generalized. Let X be any function of the elements of **H** such as occurs in the expression for F, and consider the product

$$f^{am}(f^{bn}X_{n})_{m},$$

which is equal to

$$f^{am}f^{bn}{}_{m}X_{n} + f^{am}f^{bn}X_{mn}.$$

By (2) the a and b in the first term may be interchanged, and, since X_{mn} is unaltered by interchanging m and n, the interchange of a and b in the second term also makes no change, or

$$f^{am}(f^{bn}X_{n})_{m} = f^{bm}(f^{an}X_{n})_{m}. \qquad \ldots\ldots(3)$$

Now apart from a numerical factor, each term in F is the product of two parts; the first contains only f's with a single subscript and no superscript; and the second f's with two superscripts, some of which are also differentiated. By repeated application of (3) we may make interchanges of the letters $a, b, c, \ldots$ (which do not imply summation when this second part is considered in isolation from the first) without altering the value of the part. When F is differentiated with respect to the variable h, the differentiation of the first part of any term is represented by substituting a double subscript $ah, bh, ch, \ldots$ for one of the single subscripts $a, b, c, \ldots$ Let the second part of the product be arranged with the corresponding factors f^{at}, f^{bt}, f^{ct}, $\ldots$ respectively in the leading position. Summation with respect to the common letter then leads by (1) to zero values except when $t=h$, and, whichever of the factors $f_a, f_b, f_c, \ldots$ is differentiated, the contributions to F_h are all equal. The total contribution from a term with $r+1$ factors in the first part is therefore $r+1$ times that resulting from the differentiation of only one of these factors, and this multiplier converts the numerical factor $\dfrac{1}{(r+1)!}$ to $\dfrac{1}{r!}$. The differentiation of the second part of any term is of course represented by the addition to it of the suffix h. Thus

$$\begin{aligned}F_h=f_h&-\frac{1}{2!}[2f_af_{bh}f^{ab}+f_af_bf^{ab}{}_h]\\&+\frac{1}{3!}[3f_af_bf_{ch}f^{cp}f^{ab}{}_p+f_af_bf_c(f^{cp}f^{ab}{}_p)_h]\\&-\frac{1}{4!}[4f_af_bf_cf_{dh}f^{dq}(f^{cp}f^{ab}{}_p)_q+f_af_bf_cf_d\{f^{dq}(f^{cp}f^{ab}{}_p)_q\}_h]\\&+\dots\\=f_h&-\frac{1}{2!}[2f_h+f_af_bf^{ab}{}_h]\\&+\frac{1}{3!}[3f_af_bf^{ab}{}_h+f_af_bf_c(f^{cp}f^{ab}{}_p)_h]\\&-\frac{1}{4!}[4f_af_bf_c(f^{cp}f^{ab}{}_p)_h+f_af_bf_cf_d\{f^{dq}(f^{cp}f^{ab}{}_p)_q\}_h]\\&+\dots\\=0&,\end{aligned}$$

the result to be proved.

The work described above has been carried out as part of the research programme of the National Physical Laboratory, and this paper is published by permission of the Director of the Laboratory.

REVIEWS OF BOOKS

Researches on Normal and Defective Colour Vision, by W. D. WRIGHT. Pp. xvi + 383. (London: Henry Kimpton, 1946.) 36*s*.

In this book are collected together the results of twenty years of research on colour vision by Dr. W. D. Wright and his co-workers at the Imperial College. The record includes several of the cardinal modern investigations on the subject: the determination of the colour coefficients of the spectrum colours for trichromats, anomalous trichromats and dichromats, measurements of the hue and saturation limens for all these classes and of the general colour limen for trichromats, the development of the method of binocular colour-matching for the study of colour adaptation and for the determination of the fundamental response curves of the trichromatic mechanisms, the comparison of luminosity curves in the fovea and parafovea at different brightness levels, and the quantitative study of colour and brightness matching in very small matching fields. Most of the measurements were carried out with the Wright trichromatic colorimeter, suitably modified where necessary, and an adequate account of this instrument and of the experimental technique is included. Two preliminary chapters summarize the knowledge of the visual process necessary for the proper understanding of what follows. These preliminary chapters apart, the investigation of workers outside the Imperial College group are considered only as far as is necessary for the discussion of the results of the author and his collaborators.

In re-presenting this work, together with some hitherto unpublished extensions of it, Dr. Wright has had the opportunity of pruning away the less significant material, of correcting a few erroneous conclusions and of reviewing the whole in the light of the latest information. All this is admirably done. The result is a clear exposition both of the basic experimental studies themselves and of the considered views of the author on their interpretation. As such, the book is indispensable to every research worker on vision.

It can also be strongly recommended to those concerned with the technological or medical aspects of colour.

Dr. Wright asks that his book should be regarded primarily as a record of experimental data. Much earlier work on colour vision—and even some current work—is of little value because of the inadequate specification of experimental conditions and results. It is very pleasant, therefore, to find a writer on the subject who foresees possible questions about the precise conditions of observation and takes care to provide the answers. One might wish perhaps that some of the curves showing the recovery of the eye after colour adaptation could have been plotted in terms of the fundamental stimuli (say Pitt's fundamental stimuli) instead of the instrumental primaries. Presumably this would not have altered any conclusions, but the discussion of the results would have been facilitated.

There are a few points in the exposition which are obscure or open to question. Insufficient emphasis is laid on the fact that heterochromatic brightness-matching (direct comparison) is an operation of a lower order of certainty than trichromatic colour matching. It is a pity that the notion of luminosity is not completely excluded from the original statement of the laws of trichromatic matching. Thus on page 108 the equation $x_1 = u_1 + v_1 + w_1$ and the two similar equations are true only to the extent that additivity holds good in heterochromatic brightness-matching, and they might with advantage have been omitted in explaining a fundamental experiment on colour-matching. The use of " sensation magnitude " in the analysis of discrimination data (Chap. XVI) may be meat to some but will certainly be poison to others. While Dr. Wright is probably correct in assuming that matches made by the binocular method at various times after the removal of an adapting stimulus can be extrapolated back to give some property possessed by the retina while the adapting stimulus is still on, a rather fuller discussion of this point would have been welcome. A precise definition of the term " photo-chemical sensitivity " would be interesting. A statement on p. 345 suggests that because there is little change of hue with change of wave-length in a certain spectral region, radiations in that region will approximate to a fundamental stimulus acting on one trichromatic mechanism only. Surely there is the alternative that the spectral sensitivities of the mechanisms are in a constant ratio in the region in question, and this ratio may have any value. There seem to be very few minor slips, but one or two may be noted. Page 16, position of cone maximum, text and figure disagree ; p. 29, " film colour ", not " volume colour ", is the name usually applied (e.g. by Katz) to the illuminated field of an instrument ; p. 228, figure 137, caption in error ; p. 32, Hecht used two methods for arriving at the number of quanta absorbed by the visual purple at the threshold ; in the one which seems to be referred to here, the number was calculated from the sensitivity of the eye as a whole, not vice versa.

The book is well produced, but some variations in the paper provide a good example of the technological importance of small colour differences. W. S. S.

Introduction to Electron-Optics, by V. E. COSSLETT. Pp. 272. (Oxford University Press, 1946.) 20*s*.

This is the first book to appear on this subject since the War. It is a welcome addition to the texts available to English workers, since it was written in the light of experience gained in imparting the principles of the subject to final-year university students. Such a task involves acquaintance with a wide field of work bearing on applications, some of which have been made throughout the war period, and are still in process of development.

The text is divided into two clearly defined and complementary sections—the theoretical and practical, and bears evidence of the wise discrimination exercised by the author in his choice of subject material. He has followed an established practice in presenting the properties of the electrostatic field. The methods used for solving some types of problem, including the relaxation method of Southwell, form an introduction to field plotting and ray tracing of trajectories. These, in turn, serve to introduce the focusing properties of all types of electrostatic lenses—aperture, cylinder, immersion and symmetrical—from the point of view of "geometrical optics ", which is a satisfactory approach. The chapter on magnetic focusing is especially well done, for here the author has contributed to some of the researches which he describes. The complicated subject of lens aberrations concludes the first section. These are treated individually, and the accompanying physical defects are clearly stated, but, for the full appreciation of the text, a " background " seems necessary here.

The second portion is mainly descriptive. It deals briefly with the fundamental requirements of electron devices, such as electron gun, image tube, multipliers, C.R. tube, electron microscope, diffraction, β-ray spectroscopy, magnetron, cyclotron, betatron, and velocity modulated tubes. Such an interesting review of development indicates the vast field of present-day electron-optical research ; it also makes both portions of the book blend well together. The general impression gained is that Dr. Cosslett has produced a well-balanced account of the principles and their applications.

Some parts could be expanded with profit, for example the effects of space charge on focused beams, electron distribution in beams (hardly touched on), design of practical guns arising from the latest work on " crossover " properties and the errors of deflecting fields. No doubt these will be dealt with in the next edition.

It is, nevertheless, an excellent monograph on the subject, and a distinct contribution to the literature. L. JACOB.

Antennae: *An Introduction to their Theory*, by Dr. J. AHARONI. Pp. 265. (Oxford University Press, 1946.) 25*s*.

Although aerials have formed an essential and widely used element in radio communication from its very beginning, their theory is not sufficiently known even today. This is probably the reason why, until recently, there were practically no books dealing with the theory of aerials. During the last year, however, there has been a tendency to fill this gap. After the books of Pidduck and King, the book of Dr J. Aharoni means a serious effort to present as far as possible a complete survey of our theoretical knowledge on aerials.

The book begins with a clear exposition of electromagnetic theory by means of Maxwell's equations in a form appropriate for dealing with aerial circuits, using Hertzian vector and retarded electromagnetic potentials. The chapter is illustrated by a study of forced oscillations of a sphere and a prolate spheroid. The next chapter contains the theory of magnetic and electric dipoles, Hallèn's and King's solutions of current distribution and impedances of these aerials, some considerations on mutual impedance of aerials, the theory of receiving aerials, and polar diagrams of single aerials and aerial arrays. An interesting section deals with the effect of the earth, summarizing the work of Sommerfeld, Norton, Burrows, McPetrie and others. The last chapter presents Schelkunoff's theory of a dipole based on the biconical model.

The treatment of the subject in the book is purely mathematical, and it will not be easy reading for a radio engineer.

The author's intention to give an impartial survey of existing theories " without reflecting any opinion on the relative values of different methods " is a merit of the book on the one hand, but its weakness on the other, because it leaves the reader with an impression that our knowledge of aerials is a very complete one—which is far from the truth. In fact, most of the solutions for aerials of finite thickness are given in the form of infinite series, whose convergence has not been checked. J. H. Tait, of S.R.D.E., has shown, for example, in an as yet unpublished paper, that the classical Hallèn's solution of a dipole with an infinitesimal gap is divergent, and finite results for input impedance obtained by Hallèn, as well as by others using similar methods (King, Harrison, Blake, etc.), are probably due only to the approximations introduced into the original equation.

It is to be regretted that the author does not mention some recent experimental results, some of which do not quite confirm the validity of the existing theoretical results.

In spite of this criticism, the book is a valuable contribution to the literature of the subject. B. STARNECKI.

THE PROCEEDINGS OF
THE PHYSICAL SOCIETY

VOL. 59, PART 3 1 May 1947 No. 333

THE ADIABATIC TEMPERATURE CHANGES ACCOMPANYING THE MAGNETIZATION OF COBALT IN LOW AND MODERATE FIELDS

By L. F. BATES AND A. S. EDMONDSON,
University College, Nottingham

MS. received 6 *September* 1946

ABSTRACT. The new method devised for the measurement of the small thermal changes which are associated with the step-by-step changes in the magnetization of ferromagnetic materials in fields not exceeding a few hundred oersteds has been used in the study of annealed and unannealed cobalt in the form of stout wire. The observed changes are relatively large and in striking contrast to those observed with iron and nickel. An attempt is made to explain them on the basis of modern concepts in ferromagnetism.

§ 1. INTRODUCTION

EXTENSIVE investigations of the temperature changes which occur when a ferromagnetic substance is taken through an ordinary or so-called "technical" hysteresis cycle were made by Bates and Weston (1941) in the case of nickel and several nickel-iron alloys, and the results were described in a paper, hereafter referred to as Paper I, in which references to earlier work by other experimenters may be found. The investigations were extended to specimens of Armco iron by Bates and Healey (1943), described as Paper II. The present communication deals with work on cobalt. This metal, in the form of annealed specimens of electrolytic origin, was examined by Okamura (1936) who mounted bars of cobalt 12 cm. long alternately with bars of German silver in a cylindrical frame, and arranged a system of some 31 to 47 thermocouples in series by connecting the appropriate ends of these bars with wires of copper and constantan. Such an arrangement could not be regarded as entirely satisfactory, as is borne out by the fact that the thermal changes recorded experimentally were some 10 to 20 per cent greater than those calculated from the areas of the corresponding hysteresis cycles. Moreover, in presenting his results, Okamura divided the observed thermal changes into two parts, which he termed reversible and irreversible respectively, and we find it difficult to understand the argument upon which this division is based.

The cobalt used in our work was kindly supplied by Messrs. Brandhurst & Co. Ltd. in the form of No. 12 S.W.G. wire. Its composition was : Co 98·40, Ni 0·43, Fe 0·13, CaO 0·23, Mn 0·08, C 0·19, Zn 0·01, Mg 0·11, SiO_2 0·14, S 0·02 per cent : the loss observed on heating in hydrogen was 0·24 per cent. The metal was originally cast into small ingots which were then cogged and rolled into bars from which the wire was drawn. Measurements were made on the material in the hard-drawn state, exactly as supplied,and also on wires which were annealed by heating them at 700° c. in an evacuated quartz tube for 60 minutes and thereafter allowing them to cool slowly. This annealing process is generally held to produce re-crystallization without undue increase in grain size.

A 40-cm. length of cobalt wire was mounted along the axis of the vertical water-cooled solenoid, as in Papers I and II. Adiabatic temperature changes of the wire were measured by means of twenty copper-constantan thermocouples. The " hot " junction of each couple was kept in moderately loose contact with the wire, while the " cold " junction was very close to, but thermally insulated from, the wire, except for conduction along the material of the couple. Moderately loose contact meant that the specimen was in no wise strained or prevented from changing its dimensions freely due to changes in magnetization. Each couple was joined to its own separate primary winding of insulated low-resistance copper wire wound upon a section of a mu-metal spiral core. A low-resistance secondary coil of many turns was wound upon this core and connected to a specially designed fluxmeter of high sensitivity. Electrical insulation between the " hot " junctions and the cobalt wire was not necessary but, in order to avoid instability of the fluxmeter zero, it was necessary to earth the specimen and all portions of the surrounding apparatus and, in addition, one of the leads to the moving coil of the fluxmeter was earthed.

When the temperature of the wire was rapidly changed by a small quantity ΔT, a ballistic deflection of the fluxmeter strictly proportional to ΔT took place. The whole system was normally calibrated in earlier work by suddenly applying a longitudinal force of F dynes to the wire, so causing an adiabatic fall in temperature ΔT_1 given by

$$\Delta T_1 = \frac{-\alpha TF}{J\rho SA},$$

where α is the coefficient of linear expansion of the cobalt, T its absolute temperature, J the mechanical equivalent of heat, while ρ, S and A are respectively the density, specific heat and area of cross-section of the wire. As the energy in ergs required to change the temperature of 1 c.c. of the wire by ΔT_1 is $J\rho S \,.\, \Delta T_1 = \alpha TF/A$, the density and specific heat need not be known in order to express the experimental results in the most convenient way.

Professor W. Wilson, in private conversation, has kindly pointed out to us that this method of calibration is based on the assumption that the thermodynamic conditions for a reversible change are satisfied, and these, in particular, require that the load should be applied slowly enough to ensure only very slight departure from equilibrium of the system at any time during the change. We think that our method of applying the load caused this requirement to be satisfied. It would appear, however, that equilibrium considerations have received little attention in dealing with the problems of the hysteresis cycle under alternating field conditions, as, for example, in considering the effect of changing the frequency on the rate of generation of heat per cycle for a given maximum intensity of an alternating magnetic field. Unfortunately, the loading method of calibration failed because the wire was so thin that it was impossible to mount it without bending. Hence, a sudden application of a longitudinal force produced irregular bending of the wire and gave fluxmeter deflections which were not strictly proportional to the force. We therefore assumed that in all cases Warburg's law was accurately obeyed, i.e. that the total heat liberated in the wire when it was taken through a complete hysteresis cycle was exactly equal to $\oint H\,dI$. This is only correct when eddy-current effects are negligible. Fortunately, towards the end of our measurements we were able to use a check method of calibration, devised by Mr. E. G. Harrison, which is based on the heating effect of a low-frequency alternating current passed through the wire for a known short interval of time, and which proved the above procedure to be sound. In any case, the fact that the values of the sensitivity of the system, as found from the data for the different cycles, assuming Warburg's law to hold, were in good agreement with one another showed that the method was reliable.

The necessary magnetic measurements were made by the ballistic method of Paper I. As the ratio of length to diameter of the specimen was so great, the value of the demagnetization coefficient was so small that it was not required to a high degree of accuracy, and it was therefore taken to be that for an ellipsoid of revolution with the appropriate dimensional ratio.

§ 2. EXPERIMENTAL DETAILS

The main sources of error found with this method were discussed fully in Paper I, and the same steps were taken to avoid errors due to zero drift in the fluxmeter, inadequate thermal insulation, eddy currents in the specimen and the effects of stray fields from solenoid and specimen upon the mu-metal core and upon the thermocouple leads. In particular, a 2-henry choke was connected in the solenoid circuit to reduce the rate of change of magnetization in the specimen and, consequently, the magnitude of the eddy-current heating. The latter was proved to be unimportant by the fact that the total heat generated in describing a given closed hysteresis cycle did not depend upon the number of steps or field changes in which it was done.

The fluxmeter itself was used under approximately the same sensitivity conditions as in Paper I, but calibration showed the overall sensitivity of the thermocouple-fluxmeter system to be disappointingly low, viz. about one-third that in Paper I. This was mainly because the "hot" junction contacts had been designed for specimens of larger diameter, with which greater areas of contact were possible. Attempts to improve the overall sensitivity by reducing the size of the junctions and altering the mode of attachment gave very little result.

The troublesome induction effect of unknown origin was present in magnitude rather greater than in Paper I but much less than in Paper II, and it persisted in spite of great care in arranging the thermocouple leads etc., as symmetrically as possible. As in Paper II, its effects were compensated by adjusting the compensating coil by trial prior to the recording of the data for a chosen step in the magnetizing current.

§ 3. EXPERIMENTAL RESULTS

The results for unannealed cobalt are given in figures 1 to 5. The first shows the three main hysteresis cycles used in the work. It is clear that measurements with much higher fields would have been very informative, as we barely attained even "technical" saturation, but these would have required a new solenoid of many more turns and a more complicated cooling system. Following the procedure of Papers I and II, the heat changes ΣdQ are denoted by Q; these were recorded and summed, in the cases of figures 3, 4 and 5 *a*, as the effective soleniod field was changed step by step from the stated maximum value, $-H_m$, to an equal maximum, $+H_m$, in the opposite sense. The values of Q are plotted as a function of the observed intensity of magnetization of the specimen, and, in order to economize in graph space, the values for this half-cycle only are plotted. The changes which occurred in the half-cycle from $+H_m$ to $-H_m$ would give the graph obtained by rotating the existing Q curve about the axis of ordinates and displacing it vertically

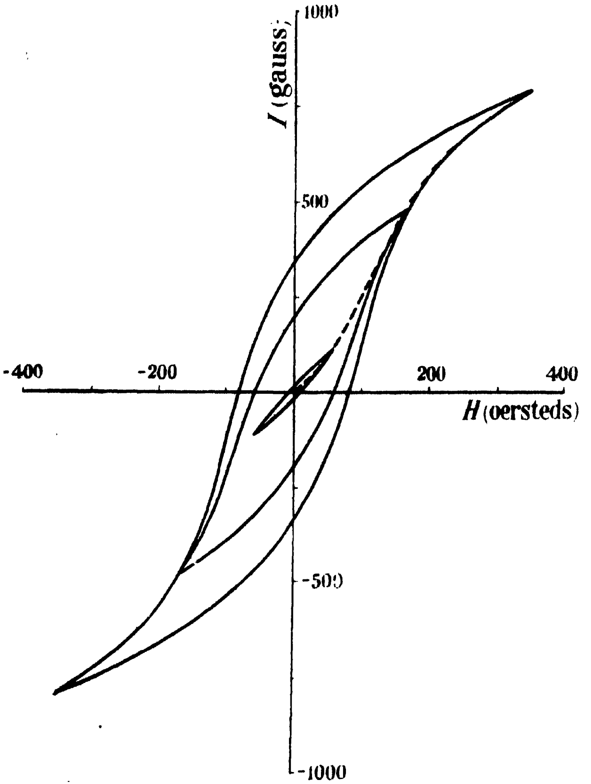


Figure 1. Hysteresis cycles for unannealed cobalt.

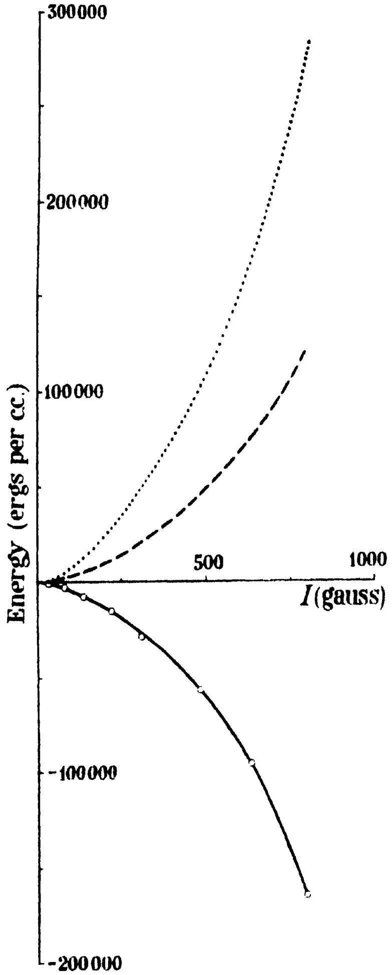


Figure 2. Virgin curves for unannealed cobalt.

—— Q, I; - - - - $\int H\,dI, I$;
...... $\int H\,dI - Q, I$.

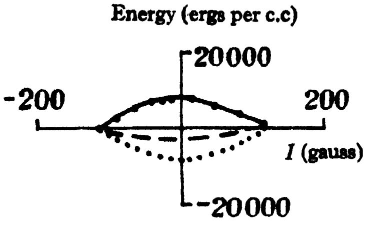


Figure 3. Unannealed cobalt. Cycle A. Maximum field 59 oersteds.

—— Q, I ; - - - - $\int H\,dI, I$;

. . . . . . $\int H\,dI - Q, I.$

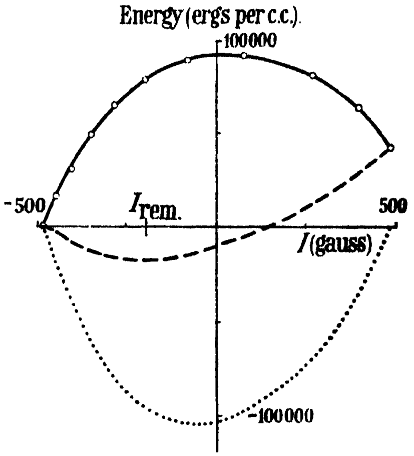


Figure 4. Unannealed cobalt. Cycle B. Maximum field 175 oersteds.

—— Q, I ; - - - - $\int H\,dI, I$;

. . . . . . $\int H\,dI - Q, I.$

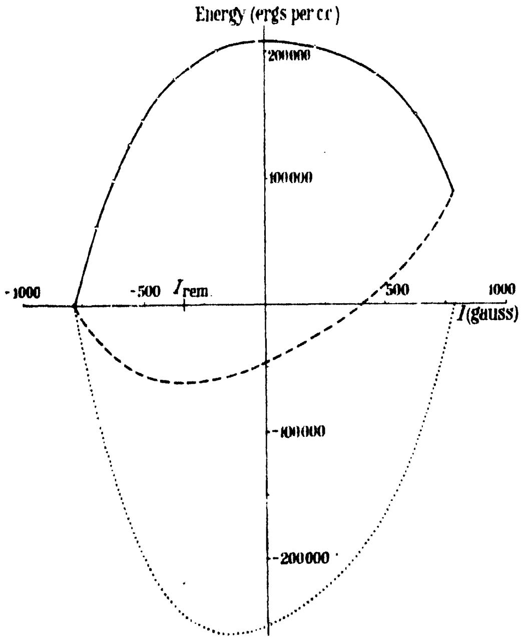


Figure 5 *a*. Unannealed cobalt. Cycle C. Maximum field 351 oersteds.

—— Q, I ; - - - - $\int H\,dI, I$;

. . . . . . $\int H\,dI - Q, I.$

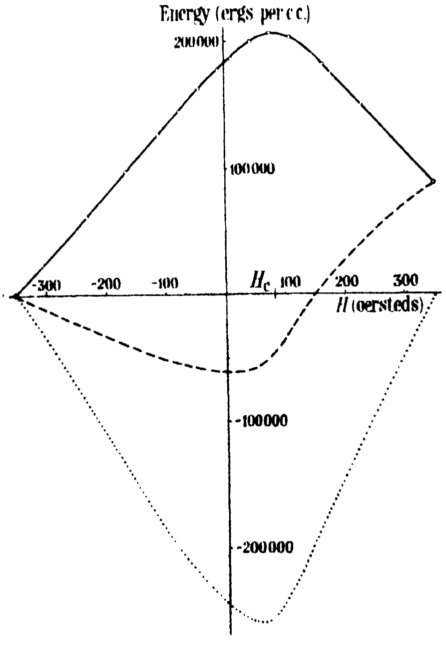


Figure 5 *b*. Unannealed cobalt. Cycle C. Maximum field 351 oersteds.

—— Q, H ; - - - - $\int H\,dI, H$;

. . . . . . $\int H\,dI - Q, H.$

until its starting point coincided with the point at which the existing curve ends. In addition, the values of $\int_{-H_m}^{H} H\,dI$ and $\int_{-H_m}^{H} H\,dI - Q$ are plotted against I.

The scales have been kept the same, wherever possible, to facilitate comparison. The relevant data are given in tables 1 to 4.

In figure 5*b* these several quantities have been plotted against H in order to bring out the differences in behaviour of unannealed iron and nickel on the one hand and of unannealed cobalt on the other. Figure 5*b* should be compared with figure 5 of Paper I and with figure 1*b* of Paper II. In the cases of unannealed iron and nickel there is always an initial cooling followed by a heating such that the

Table 1. Unannealed cobalt—Cycle A

Step	H (oersteds)	I (gauss)	$\int H\,dI$ (ergs/c.c.)	Q (ergs/c.c.)	E (ergs/c.c.)
0	−59·1	−114·0	0	0	0
1	−45·3	− 91·7	−1,130	+2,690	− 3,820
2	−29·6	− 66·7	−2,060	+5,220	− 7,280
3	−15·0	− 41·1	−2,640	+7,040	− 9,680
4	− 6·0	− 25·3	−2,800	+7,200	−10,000
5	− 0·5	− 15·3	−2,840	+7,360	−10,200
6	+ 0·5	− 13·6	−2,840	+7,360	−10,200
7	+15·1	+ 14·6	−2,620	+6,760	− 9,400
8	+29·7	+ 46·1	−1,910	+5,600	− 7,510
9	+45·3	+ 81·3	− 580	+3,740	− 4,320
10	+59·1	+114·0	+ 990	+ 990	0

Table 2. Unannealed cobalt—Cycle B

Step	H (oersteds)	I (gauss)	$\int H\,dI$ (ergs/c.c.)	Q (ergs/c.c.)	E (ergs/c.c.)
0	−175·0	−484	0	0	0
1	−143·0	−451	− 1,500	+16,400	− 17,900
2	−109·0	−407	− 8,060	+32,700	− 40,800
3	− 73·6	−351	−12,200	+49,200	− 61,400
4	− 37·7	−286	−15,900	+64,800	− 80,700
5	− 0·5	−202	−17,500	+78,500	− 96,000
6	+ 1·7	−197	−17,500	+78,500	− 96,000
7	+ 38·8	− 82	−15,100	+89,100	−104,200
8	+ 74·5	+ 76	+ 5,980	+91,600	− 97,600
9	+110·0	+266	+11,500	+80,500	− 69,600
10	+144·0	+396	+27,900	+63,400	− 35,500
11	+175·0	+484	+42,000	+42,000	0

Table 3. Unannealed cobalt—Cycle C

Step	H (oersteds)	I (gauss)	$\int HdI$ (ergs/c.c.)	Q (ergs/c.c.)	E (ergs/c.c.)
0	−351·0	−790	0	0	0
1	−231·0	−699	−26,800	+ 61,300	− 88,100
2	−166·0	−628	−40,500	+ 98,000	−138,000
3	−111·0	−557	−50,600	+127,000	−178,000
4	− 76·5	−503	−55,800	+145 000	−201,000
5	− 43·2	−438	−59,600	+163,000	−223,000
6	− 10·5	−368	−61,600	+178,000	−240,000
7	+ 12·5	−307	−61,600	+187,000	−249,000
8	+ 45·1	−197	−58,400	+200,000	−258,000
9	+ 78·1	− 25	−47,500	+207,000	−254,000
10	+112·0	+218	−24,300	+203,000	−227,000
11	+166·0	+471	+10,500	+182,000	−172,000
12	+231·0	+631	+42,000	+150,000	−108,000
13	+351·0	+790	+87,600	+ 87,600	0

Table 4. Unannealed cobalt—Virgin curve

Step	H (oersteds)	I (gauss)	$\int HdI$ (ergs/c.c.)	Q (ergs/c.c.)	E (ergs/c.c.)
0	0·0	0·0	0	0	0
1	+ 0·4	+ 0·6	0	0	0
2	+ 22·9	+ 36·4	+ 450	− 800	+ 1,250
3	+ 45·1	+ 82·7	+ 2,040	− 2,700	+ 4,740
4	+ 67·6	+138·0	+ 5,180	− 7,800	+ 13,000
5	+ 90·8	+220·0	+ 11,700	− 15,700	+ 27,400
6	+115·0	+310·0	+ 21,100	− 29,000	+ 50,100
7	+168·0	+485·0	+ 45,700	− 57,000	+103,000
8	+232·0	+634·0	+ 75,400	− 96,000	+171,000
9	+351·0	+795·0	+121,000	−165,000	+286,000

maximum fall of temperature occurs in the interval of field change from $-H_m$ to $-H_c$. With cobalt there is always an initial warming, such that the maximum rise of temperature takes place in the field change from $-H_m$ to $+H_c$. Here, $+H_c$ means the coercive field actually required to reduce the magnetization to zero; heating is observed in all cases, of course, as the field is changed from $-H_c$ to $+H_c$, in accord with the view that, between these field limits, changes in magnetization take place almost entirely by irreversible 180-degree reversals in the domains.

The results for annealed cobalt are given in figures 6 to 10; the relevant data for figure 10 are given in tables 5 to 8. Incidentally, the results of figure 9 were

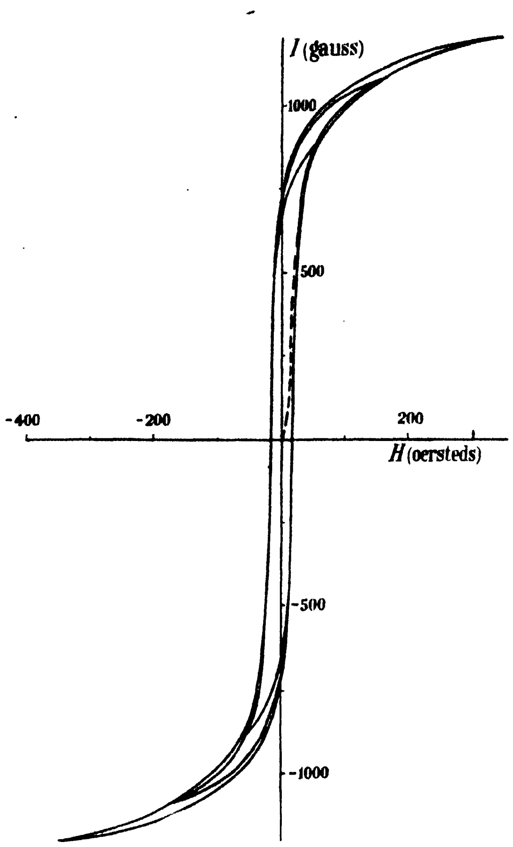


Figure 6. Hysteresis cycles for annealed cobalt.

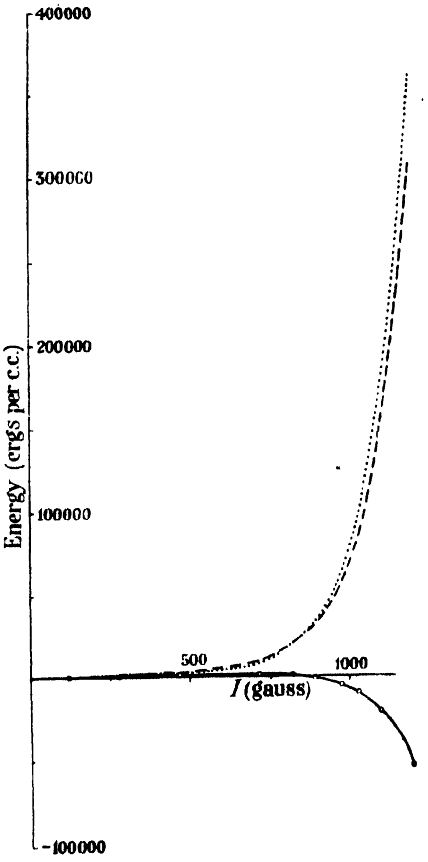


Figure 7. Virgin curves for annealed cobalt.

—— Q, I ; - - - - $\int H\,dI, I$;

…… $\int H\,dI - Q, I$.

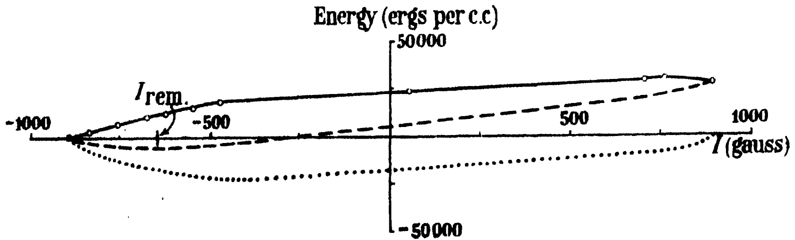


Figure 8. Annealed cobalt. Cycle A. Maximum field 57 oersteds.

—— Q, I ; - - - - $\int H\,dI, I$; …… $\int H\,dI - Q, I$.

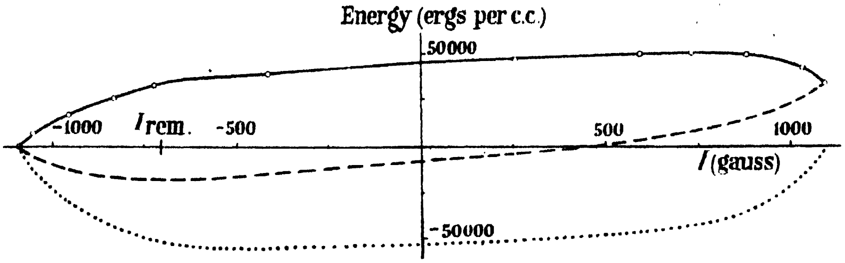


Figure 9. Annealed cobalt. Cycle B. Maximum field 173 oersteds.

—— Q, I ; - - - - $\int H dI. I$; . . . . . $\int H dI - Q, I$.

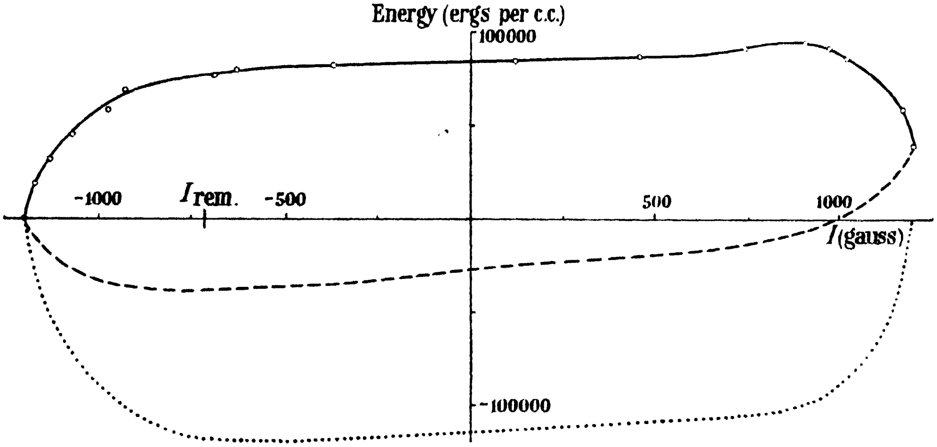


Figure 10. Annealed cobalt. Cycle C. Maximum field 350 oersteds.

—— Q, I ; - - - - $\int HdI, I$; . . . . . $\int HdI - Q, I$.

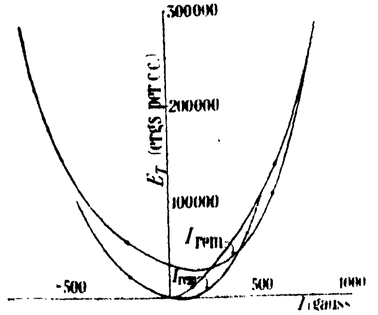


Figure 11 *a*. Curves of E_T against I for unannealed cobalt.

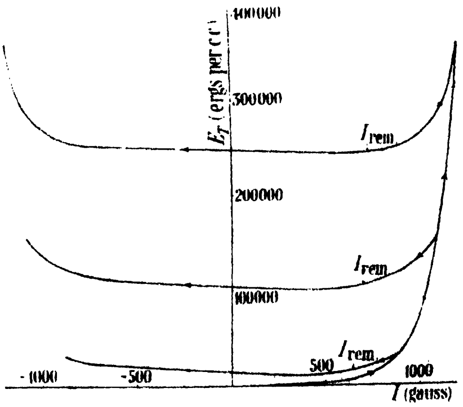


Figure 11 *b*. Curves of E_T against I for annealed cobalt.

fully confirmed by Mr. E. G. Harrison, who completely dismantled the specimen, overhauled the thermocouple system, inserted a larger choke in the solenoid circuit and changed its cooling arrangements. He also obtained the whole of the results of figures 7 and 10. The changes in coercivity produced by annealing are remarkable. Qualitatively, they might be explained in terms of the magnetization of single crystals of cobalt. Kaya (1928) showed that in cobalt, which possesses hexagonal structure, the (0001) direction is the direction of easy magnetization while the (1010) direction, in the plane of the hexagon, is a direction of difficult magnetization, requiring fields of some ten thousand oersteds to produce technical saturation. If in the hard-drawn state many of the crystals in the polycrystalline wire tended to set with their directions of easy magnetization perpendicular to the axis of the wire, it would be difficult to produce saturation parallel to that axis. If, however, the effect of annealing is to reduce the amount of internal strain in the

Table 5. Annealed cobalt—Cycle A

Step	H (oersteds)	I (gauss)	$\int H\,dI$ (ergs/c.c.)	Q (ergs/c.c.)	E (ergs/c.c.)
0	−56·5	−891	0	0	0
1	−34·1	−838	− 2,420	+ 3,020	− 5,440
2	−14·1	−758	− 4,160	+ 7,340	−11,500
3	− 3·2	−678	− 4,690	+10,900	−15,600
4	+ 2·0	−625	− 4,720	+13,000	−17,700
5	+ 7·0	−550	− 4,440	+15,500	−19,900
6	+10·1	−473	− 3,530	+18,400	−21,900
7	+16·2	+ 52	+ 6,650	+24,100	−17,400
8	+30·0	+704	+19,900	+29,200	− 9,300
9	+34·3	+760	+21,800	+30,000	− 8,200
10	+56·5	+891	+27,600	+27,600	0

Table 6. Annealed cobalt—Cycle B

Step	H (oersteds)	I (gauss)	$\int H\,dI$ (ergs/c.c.)	Q (ergs/c.c.)	E (ergs/c.c.)
0	−173·0	−1,090	0	0	0
1	−116·0	−1,050	− 5,210	+ 7,300	−12,500
2	− 53·9	− 956	−12,600	+17,200	−29,800
3	− 17·7	− 835	−16,700	+25,900	−42,600
4	− 2·2	− 727	−17,700	+32,800	−50,500
5	+ 13·4	− 416	−15,200	+39,200	−54,400
6	+ 19·4	+ 255	− 4,260	+46,600	−50,900
7	+ 26·0	+ 593	+ 3,040	+49,600	−46,600
8	+ 34·0	+ 733	+ 7,170	+50,200	−43,000
9	+ 54·1	+ 882	+13,600	+49,000	−35,400
10	+116·0	+1,030	+25,800	+41,500	−15,700
11	+173·0	+1,090	+33,300	+33,300	0

Table 7. Annealed cobalt—Cycle C

Step	*H* (oersteds)	*I* (gauss)	$\int H\,dI$ (ergs/c.c.)	*Q* (ergs/c.c.)	*E* (ergs/c.c.)
0	−350·0	−1,120	0	0	0
1	−252·0	−1,170	− 9,970	+18.900	− 28,900
2	−176·0	−1,130	−19,200	+31,800	− 51,000
3	−116·0	−1,070	−26,700	+44,800	− 71,500
4	− 54·7	− 976	−34,400	+58,000	− 92,400
5	− 37·5	− 930	−36,400	+68,500	−105,000
6	+ 2·3	− 692	−38,800	+76,200	−115,000
7	+ 6·8	− 632	−38,600	+79,200	−118,000
8	+ 13·7	− 371	−36,100	+81,800	−120,000
9	+ 18·1	+ 122	−24,700	+84,400	−109,000
10	+ 23·0	+ 458	−20,300	+86,600	−107,000
11	+ 36·2	+ 746	−14,200	+91,400	−106,000
12	+ 65·9	+ 909	− 6,220	+94,300	−100,000
13	+ 95·5	+ 973	− 120	+91,700	− 91,800
14	+116·0	+1,020	+ 3,260	+86,100	− 82,800
15	+265·0	+1,170	+29,300	+58,600	− 29,300
16	+350·0	+1,200	+38,700	+38,700	0

Table 8. Annealed cobalt—Virgin curve

Step	*H* (oersteds)	*I* (gauss)	$\int H\,dI$ (ergs/c.c.)	*Q* (ergs/c.c.)	*E* (ergs/c.c.)
0	0	0	0	0	0
1	+ 11·2	+ 121	+ 506	+ 377	+ 129
2	+ 15·3	+ 279	+ 1,270	+ 659	+ 611
3	+ 20·3	+ 468	+ 3,140	+ 800	+ 2,340
4	+ 32·0	+ 718	+ 10,200	+ 942	+ 9,290
5	+ 45·4	+ 824	+ 20,600	+ 612	+ 20,000
6	+ 58·5	+ 893	+ 31,800	− 1,080	+ 32,900
7	+ 85·3	+ 977	+ 57,100	− 5,220	+ 62,300
8	+ 114·0	+1,030	+ 85,400	−10,000	+ 95,400
9	+ 174·0	+1,100	+149,000	−21,000	+170,000
10	+ 349·0	+1,200	+307,000	−53,600	+361,000

wire and also permit recrystallization with the directions of easy magnetization more favourably directed, there would result an increase in remanence and a decrease in coercivity. However, from what follows, it appears that the result of annealing is to produce big changes in crystal structure instead of mere orientation effects.

The arched shape of the Q, I curves appears to be a distinctive feature of unannealed cobalt. If we compare figures 3, 4 and 5 *a* above with figures 13 and 14 of Paper I for annealed nickel, some slight resemblance is seen; it is somewhat intensified by annealing the cobalt, when the arches become much flattened

Indeed, the graphs of figures 8 to 10 between $-I_r$ and $+I_r$, i.e. between the two extreme values of the retentivity, are linear within the limits of experimental error. The magnitudes of the heat changes are considerably greater than those observed in corresponding cycles for iron and nickel.

In table 9 an attempt has been made to summarize the more distinctive features of the Q, I curves for the three ferromagnetic metals, and we see that, on the whole, cobalt behaves in a manner contrary to that of iron and nickel.

Table 9

	Field change	Heat changes in closed hysteresis cycles		
		Iron	Nickel	Cobalt
Unannealed metals	$-H_m$ to $-H_c$	Cooling	Cooling	Heating
	$-H_c$ to $+H_c$	Heating	Heating	Heating
	$+H_c$ to $+H_m$	Heating	Heating	Cooling
Annealed metals	$-H_m$ to $-H_c$	Cooling	Complicated	Heating
	$-H_c$ to $+H_c$	Cooling	Cooling	Heating
	$+4H_c$* to $+H_m$	Heating	Heating	Cooling

* $+4H_c$ is usually taken as a reasonable value of H at which technical saturation may be assumed complete ; here, it is given merely as an indication that complicated phenomena around $H = +H_c$ are excluded.

Unfortunately, Bates and Healey were unable to obtain the virgin Q, I curve for iron; arrangements are now being made to do so, but success is doubtful owing to the special difficulties peculiar to iron. Consequently, only the results for nickel may be compared with those for cobalt. Noting the change of abscissae, figure 2 above may be compared with figure 6 of Paper I, showing that unannealed virgin cobalt cools as it is magnetized while unannealed nickel warms, and that there is in both cases the suggestion of a " knee " in the Q, H curve (not reproduced here).

In addition to the figures published in Paper I, Bates and Weston (Weston, 1940) made many measurements of the Q, I curves for annealed virgin nickel, finding that such curves rise very sharply from the origin to reach a maximum at a magnetization equal to approximately 70 to 80 per cent of the saturation value, when they drop steeply to cross the axis of abscissae to the side where cooling is represented. The contrast with figure 7 above is striking. Mr. E. G. Harrison has extended the range of the latter curve and finds that strong cooling is still shown as even higher values of I are reached. The question therefore arises as to the stage in magnetization at which the cobalt and the nickel curves turn upwards, for we know that every ferromagnetic metal exhibits the normal magnetocaloric effect, namely a heating directly proportional to the intensity of the strong magnetic field in which it is placed suddenly. As a matter of fact, Bates and Weston found that one specimen of nickel, specimen 1A, showed an upward turn at $H = 100$ oersteds and $I = 420$ gauss approximately. Incidentally, as this specimen of nickel was subjected to increasing longitudinal stress, the Q, I curve moved to lie in its entirety above the I axis, after which, with increasing tension, the curve approached closer and closer to that axis.

§4. DISCUSSION OF RESULTS

The curves in the preceding section have been plotted in the same manner as those of Papers I and II in order to facilitate comparison, but much is gained, for example, by plotting all the Q, I curves for annealed cobalt on the same graph. One then appreciates the flat initial portion of the virgin curve and the low arches of the succeeding half-cycle curves. In like manner, in the case of unannealed cobalt, one sees the steep descent of the virgin curve and the strongly arched curves of the succeeding half-cycles. Similarly the $\int H\,dI, I$ curves can be treated to emphasize the great differences between the two kinds of cobalt. However, it is more profitable to include here the curves of $\int H\,dI - Q$ against I; these are shown in figures 11 *a* and 11 *b*, in which (to avoid confusion with preceding figures) the arrows on the curves indicate the direction in which the magnetization was changed. The most prominent feature with annealed cobalt is the magnitude of the changes in $\int H\,dI - Q$, this quantity being denoted by E_T in view of the theoretical discussion given later, as I is changed from zero to $I_{\max}$ along any chosen half-cycle, compared with the change in E_T between the same limits on the virgin curve; the former is much smaller. The opposite holds for annealed nickel, where the central portions of the E_T, I curves fall below the origin.

Following the treatment given in Papers I and II, we write

$$H\,dI = \left(\frac{\partial E}{\partial I}\right)_T dI + \left(\frac{\partial E}{\partial T}\right)_I dT$$

or

$$\left(\frac{\partial E}{\partial I}\right)_T dI = H\,dI - dQ,$$

where dQ is the change in energy which appears as heat and is measured directly in our experiments, while $(\partial E/\partial I)_T\,dI$ represents the change in internal energy which is non-thermal in character. The second equation holds independently of whether the change is thermodynamically reversible or not. Hence the importance of the $\int H\,dI - Q, I$ curve lies in the fact that the slope at any point on it gives the value of (dE_T/dI), i.e. $(\partial E/\partial I)_T$ for the corresponding point on the I, H curve.

Now, the magnetization of a substance, starting from the virgin unmagnetized state, proceeds mainly by virtue of magnetically reversible 90-degree displacement followed by irreversible 180-degree displacements located chiefly on the steeper parts of the hysteresis curve. On the application of a sufficiently strong field, magnetically reversible rotations of the domain vectors set in on the upper portions of the curve beyond the "knee", and these in turn are followed by the changes in the intrinsic intensity of magnetization of the domains themselves as the applied field becomes so great that the effects of hysteresis fade into insignificance and the well-known magnetocaloric effect appears.

Starting with the material magnetized to technical saturation, a reduction in the field at first permits the magnetically reversible vector rotations to take place, followed at lower fields by 90-degree boundary displacements which are mainly reversible. In the region $I = I_r$ the curve becomes steeper and, in the main, irreversible 180-degree boundary displacements predominate. Beyond $H = +H_c$ the substance is remagnetized in the opposite sense, and the more important changes are essentially the same as for the steeper outer portions of the virgin curve. Of course, the above division of processes is approximate only, since reversible and irreversible changes must occur simultaneously in all parts of the cycle.

Bates and Weston found that $(\partial E/\partial I)_T$ was zero when $I = I_r$, but this result is definitely not applicable to the case of unannealed cobalt, and only when the metal has been exposed to a high field is there indication that it holds in the case of annealed cobalt. They also found that, in general $(\partial E/\partial I)_T$, was negative in the closed cycle whenever the main magnetic processes involved were reversible, and positive when these were irreversible. The fields used in the present work were manifestly too low to produce marked reversible rotations of the domain vectors, so we find that there is no change in the sign of $(\partial E/\partial I)_T$ when I is near its maximum value.

Cobalt is peculiar in that the cast material, which, presumably, has been quenched, is frequently in the form of face-centred cubic crystals, while cobalt annealed *in vacuo* and cooled slowly possesses hexagonal crystal structure, although a purely hexagonal specimen is most unlikely. Consequently, the magnetostriction properties of these two kinds of material are markedly different. Thus cast material, according to Nagaoka and Honda (1902) behaves in the reverse manner to iron, i.e. in weak longitudinal fields the material shortens, but, as its magnetization is increased, i.e. in fields of about 800 oersteds and upwards, it lengthens. On the other hand, annealed cobalt behaves very much like nickel, i.e. it shortens on magnetization in longitudinal fields of all intensities. The graph of dl/l against H for the two materials intersect when H is about 400 oersteds.

We might therefore expect the virgin Q, H curves for the two materials to be markedly dissimilar, particularly in the region of low fields. In practice, however, very little thermal change at all occurs until fields of over 50 oersteds are reached, and both for annealed and unannealed cobalt there is practically a linear fall of temperature between 100 and 400 oersteds, the rate of fall for the unannealed being three times as great as for the annealed. Unfortunately, as the magnetostriction of cobalt is not isotropic, we cannot base reliable deductions on the differences in the virgin curves with our present meagre knowledge.

One important extension of the work remains to be made. Cobalt forms interesting alloys with copper containing up to 8 per cent of the latter metal. These alloys have a face-centred cube structure. It is proposed to obtain some of these alloys and to make measurements with them on the above lines, in order to try to find how much the differences between the behaviour of cobalt, on one hand, and iron and nickel, on the other, may be attributed to differences in crystal structure.

§ 5. ACKNOWLEDGMENTS

Much of the apparatus used in this research was purchased from a grant made to L.F.B. by the Government Grant Committee of the Royal Society. Our thanks are due to Mr. L. H. Goris of Messrs. Brandhurst & Co. Ltd. for the supply

of cobalt. We record our warm thanks to Mr. E. G. Harrison for allowing us to quote data obtained by him.

REFERENCES

Bates, L. F. and Weston, J. C., 1941. *Proc. Phys. Soc.*, **53**, 5.
Bates, L. F. and Healey, D. R., 1943. *Proc. Phys. Soc.*, **55**, 189.
Kaya, S., 1928. *Sci. Rep. Tôhoku Univ.*, **17**, 1157.
Nagaoka, B. H. and Honda, K., 1902. *Phil. Mag.*, **4**, 45.
Okamura, T., 1936. *Sci. Rep. Tôhoku Univ.*, **24**, 744.
Weston, J. C., 1940. Thesis (Ph.D. London).

Note added 12 *November* 1946

In view of the way in which crystal structure and magnetostriction must play a part in determining the variation of E_T with I, it is essential to get as much information as possible about these factors. We have been very fortunate in obtaining an x-ray examination of the two specimens of annealed and unannealed cobalt which was kindly made for us by Dr. H. Lipson of the College of Technology, Manchester. He reported that he took photographs in an ordinary single-crystal camera, the specimens being offset so that the rays hit only one edge. For this reason the specimen could not be completely rotated and was oscillated through an angle of 15 degrees.

Dr. Lipson found that the hard-drawn specimen was mainly hexagonal, but there was a slight amount of cubic phase present. Some of the lines due to the hexagonal form were broadened somewhat, as found by Dr. Lipson and Miss Edwards in their work on cobalt (*J. Inst. Metals*, **69**, 177 (1943)). There was evidence of slight preferred orientation in the (0002) reflexion, but this was not investigated in detail. The whole pattern was rather diffuse, owing, presumably, to residual stresses.

The annealed specimen showed rather better line definition and there was no evidence of preferred orientation. There was possibly a little more cubic phase present, but it was not very definitely indicated.

We therefore conclude that the specimens are nothing like as diverse in crystal structure as one might anticipate from their history or from the results reported in our communication.

INVESTIGATIONS ON ABSORPTION HYGROMETERS AT LOW TEMPERATURES

By E. GLUECKAUF,
Durham Colleges in the University of Durham

MS. received 30 *October* 1946

ABSTRACT. A method has been developed for the calibration of hygrometers and for measuring their response at low temperatures. Investigations have been made on gold-beater's skin as a hygrometric element, and on the electrolytic hygrometer described by Dunmore. Improvements concerning the response of the latter are suggested. A new optical hygrometer, based on the interference of light reflected by thin hygroscopic films, is described in detail ; its advantages are temperature-independent calibration and satisfactory response down to − 60° c.

---

§ 1. INTRODUCTION

Until about 1939, humidity measurements at high altitude were predominantly based on the hair hygrometer, which in this country was employed in the form of the Dines Meteorograph. A detailed study of the behaviour of hair at low temperatures (Glückauf, 1944) permitted the interpretation of high-altitude soundings and the unsuspectedly low humidities in the stratosphere calculated in this way (Glückauf, 1945 a) have since been confirmed by the much superior direct method of Dobson, Brewer and Cwilong (1946).

Since 1939 a number of new hygrometric elements have come into use, in particular the electrolytic hygrometer by Dunmore, used in U.S.A. (Dunmore, 1939; Diamond *et al.*, 1940), and the gold-beater's skin hygrometer which was being used in this country for radio-sondes. It was therefore desirable to investigate in some detail the working conditions of these absorption hygrometers at low temperatures, as had been done in the case of hair.

To do so, it was first necessary to develop a technique of controlling relative humidity below − 15° c., the lack of which was generally admitted to be a handicap in the development and calibration of hygrometric instruments (see Diamond *et al.*, 1940, pp. 358 and 362).

§ 2. APPARATUS

The principal features of the testing apparatus are shown in figure 1.

Air from a compressor, after passing over solid caustic soda, was dried by freezing out in a copper coil immersed in a bath of trichlorethylene cooled with solid CO_2. Any remaining moisture was removed by further drying with silica gel at a pressure of 130 atm. in the steel cylinder A. (More economical methods for the production of completely dry air could have been devised, e.g. passing the compressed air through a liquefier with inefficient heat exchanger, so that the non-liquefied dried air escapes at a temperature of about − 60° c. This would have obviated all the other apparatus and at the same time would have supplied completely dry air at a low temperature. But such apparatus could not be built under war-time conditions.) After passing through a reducing valve, the completely dried air is divided into two air currents of known mass ratio by means of the distribution stopcock B. The two air currents are cooled down in two coaxial coils immersed in a low-temperature bath, and one of them is saturated (with respect to ice) at the bath temperature. They are then reunited in the testing chamber and give an air current of well defined relative humidity which is identical with the percentage of air passed through the saturation coil.

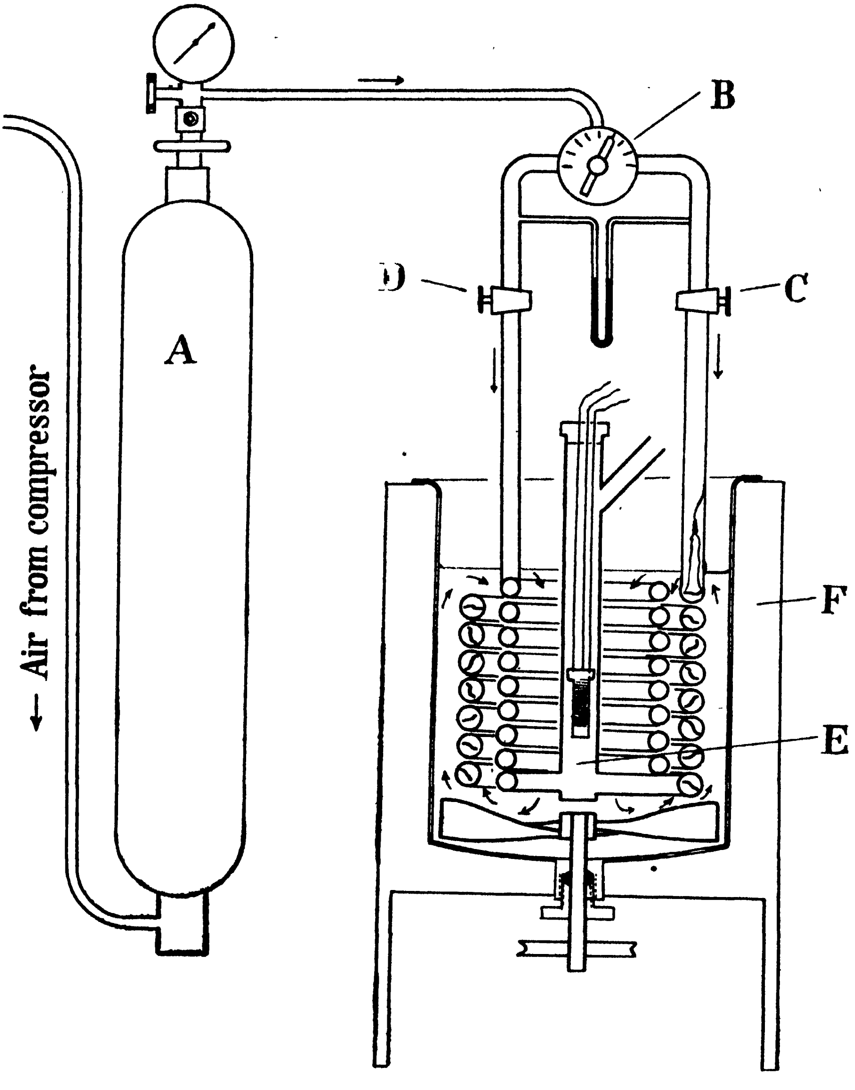


Figure 1. Apparatus for testing hygrometers at low temperatures.

A. Drying cylinder filled with silica gel.
B. Distribution stopcock.
$\left.\begin{array}{l}\text{C.}\\\text{D.}\end{array}\right\}$ Stopcocks for pressure equalization.
E. Testing chamber for hygrometric elements.
F. Lagging.

The distribution stopcock is shown in figure 2. The conical part contains nine equal holes of about 1 mm. diameter, placed at 22·5° to each other, and leading into the centre of the stopcock. If in position, one of the holes was always covered, leaving eight holes through which the air could pass. By adjusting the position of the stopcock, nine different air-flow ratios could be obtained without altering the total amount of air passing. In this way the following nominal relative humidities (with respect to ice) were obtained: 0, 12·5, 25, 37·5, 50, 62·5, 75, 87·5, 100%, which cover a range sufficient for calibrating purposes. (Due to constructional inaccuracies, the actual flow ratios, as determined by separate flow measurements in both lines, differed somewhat from the nominal ones.)

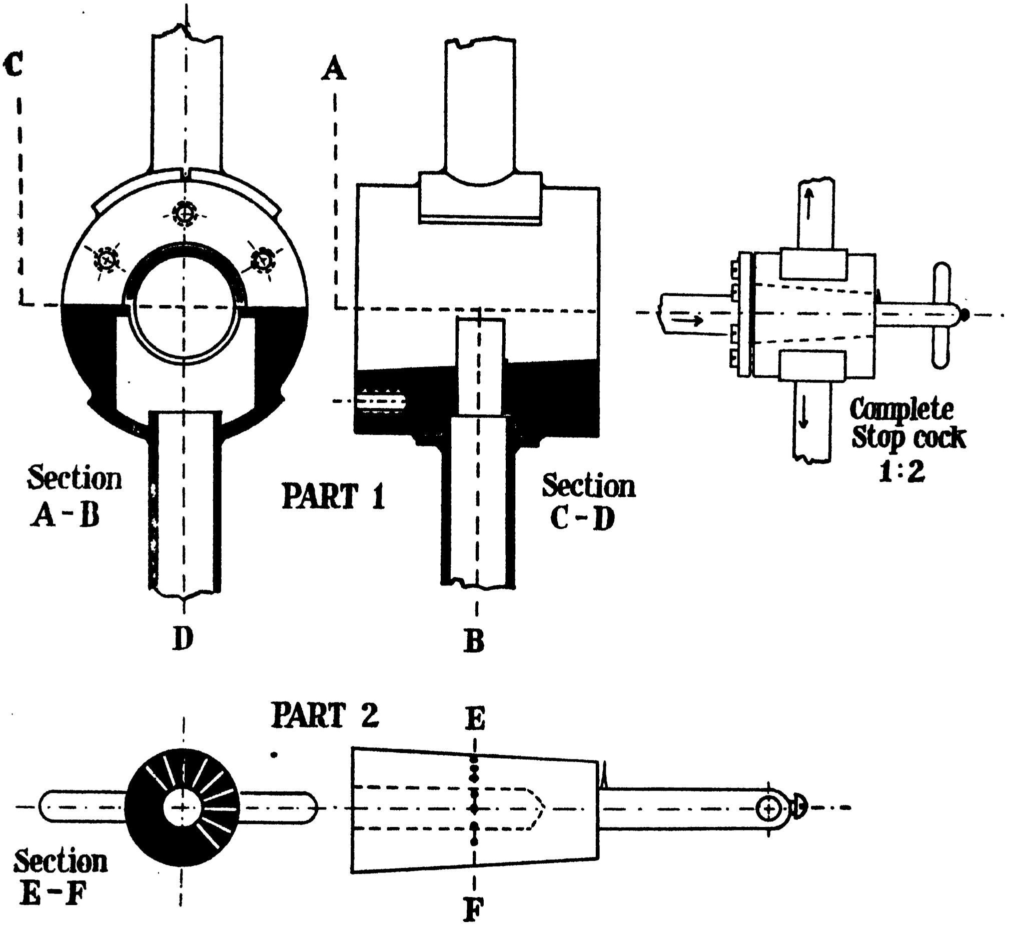


Figure 2. Distribution stopcock for dividing air stream according to well defined ratio. (1) Barrel, (2) key. Complete stopcock in reduced size.

In some earlier experiments a different arrangement using several glass stopcocks and calibrated capillaries was used which was based on the same principle but gave a greater number of combinations. It could not, however, be used for large quantities of air on account of the much higher pressures involved.

To guarantee the correct working of the distribution stopcock it was necessary to ensure equal pressure on the two outlet sides. Friction in the coils of the cooling tank caused considerable pressure differences, if the quantities of air passing through the two coils were very different (e.g. in the cases of 12·5 and 87·5% ralative humidity). To avoid errors of this kind additional flow resistance could be provided by two valves C and D, one of which was adjusted until the mercury manometer showed no difference of pressure. The arrangement made it possible to produce large well defined changes of humidity in the testing chamber within seconds.

The two air currents passing through the valves C and D were cooled down to the temperature of the bath by passing through two helical coils of tubing of ⅝ in. and ½ in. diameter respectively. The wider coil had a ribbon of cotton bandage held at the ends by copper wire and stretched through nearly the entire length of the tubing (about 15 feet). The cotton was occasionally moistened with distilled water; this arrangement proved sufficient to saturate the air completely, even at the highest speeds, with respect to ice at temperatures below −30°. The method described proved superior to the freezing out of surplus water vapour from moist air, as it avoided the formation of snow inside the tube, which at the high air speeds used would have been driven right through the apparatus. In the case of higher temperatures, however, the "wet" air stream was first moistened by passing through a wash-bottle filled with a suitable solution, so as to prevent the wet ribbon losing its water too quickly.

The two coils, for dry and saturated air respectively, were connected eccentrically to the bottom of a tube of 1 inch internal diameter, which served as testing chamber (E).

The eccentric connection is essential, as otherwise the two air streams do not mix efficiently, especially at very high and very low humidities, where the velocities in the two feeding tubes differ greatly.

The low-temperature bath was filled with trichlorethylene. As the bath had a fairly large capacity, a good temperature constancy was obtained by continuously adding small pieces of solid CO_2.

To obtain a uniform temperature, the liquid was vigorously agitated by a motor-driven stirrer, the shaft of which passed through a stuffing gland in the bottom of the tank. The blades were bent in such a way as to produce a very fast circulation of liquid in the direction indicated in figure 1. The bath container was held in a wooden frame and was only moderately insulated (F), as the transfer of heat through the insulation was negligible compared with the heat introduced by the passing air.

The units to be tested were either inserted directly into the testing chamber (as shown in figure 1, with an electrolytic hygrometer) or, if very high air velocities were required, they were held in ebonite holders which narrowed down the cross-section of the air flow. In the experiments with air velocities of about 100 miles/hour (45 m./second) the cross-section of air flow was thus narrowed down to 1 sq. cm.

§ 3. THE GOLDBEATER'S-SKIN HYGROMETER

(i) *Calibration curve of the G.B.S. at low temperatures*

Calibration curves have been taken for relative humidities from 0 to 100% at temperatures ranging from +18° c. to −65° c. These show that, at temperatures below zero, the "G.B.S." indicates relative humidity with respect to supercooled water, similar to the hair hygrometer. Table 1 gives the readings of the G.B.S. hygrometer at −21° c. in terms of the calibration at room temperature.

Table 1

% R.H. of air used relative to ice	% R.H. indicated according to room-temperature calibration	% R.H. of air used, relative to supercooled water
100	81, 78·5, 79	81·0
85	70	69·8
67·5	54, 55	54·6
57·5	46	46·6
52·5	41·5	42·5
42·5	34, 33	34·4
32·5	26, 27·5, 27	26·3
15	11·5, 12·5	12·1

In another series of experiments the relative humidity indicated by the G.B.S. was measured in air fully saturated over ice. The figures are given in table 2. Both tables show clearly that calibrations at room temperature can be applied without corrections to all temperatures, provided that the relative humidity so obtained is related to supercooled water and not to ice. The term relative humidity", or "R.H.", is therefore used in future only with respect to supercooled water, unless the reference to ice is specially stated.

Table 2

Temperature (°c.)	% R.H. indicated according to room-temperature calibration	% R.H. of air used, relative to water
+ 0·8	100	100
− 2·3	97·5	97·9
− 6·5	93·5	94·0
−10·5	90·5	90·5
−15·5	85·1	86·1
−19·2	82·6	82·9
−22·0	80·4	80·7
−25·6	77·5	77·7
−30·0	73·7	74·2
−33·5	72·5	71·6
−36·5	70·0	69·6
−39·5	67·1	67·5
−42·5	65·7	65·6
−65	53	52·6

(ii) *Hysteresis effects after exposure to low humidities*

Like the hair hygrometer, the G.B.S. hygrometer also suffers from a hysteresis effect after exposure to relative humidities lower than 30% with respect to water (see figure 3). This hysteresis effect takes place also at low temperatures; curves similar to those of figure 3, some of them almost identical, have been observed at −21° c. and −29° c. The effect of exposure to low humidities disappears only at humidities higher than 70% R.H. with respect to water, and this also occurs at −29° c. As, below −40° c., ice saturation corresponds to less than 70% R.H., the hysteresis effects are no longer relieved by occasional high relative humidities, and considerable uncertainty as to the interpretation of the readings must be accepted at these low temperatures.

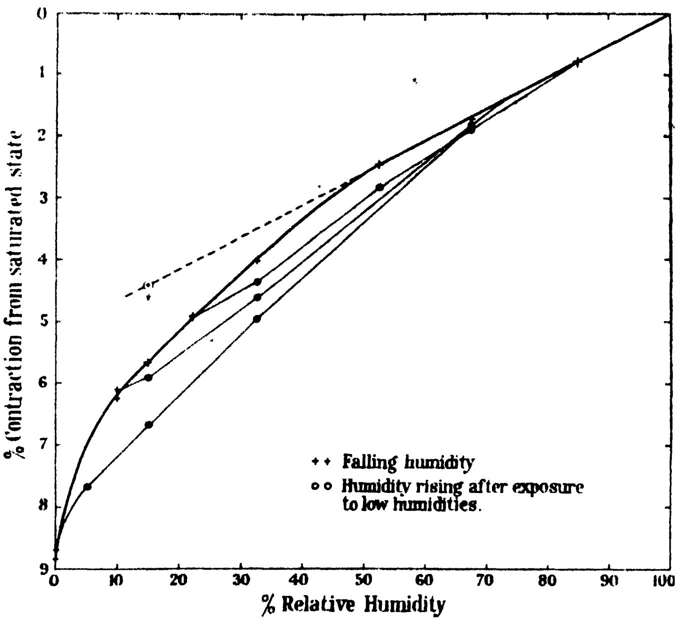


Figure 3. Calibration curve of gold-beater's skin at 12° c. (showing hysteresis curves).

Exposure at −40° to relative humidities as high as 35% R.H. (ice) may cause hysteresis errors in subsequent readings as high as 10% R.H. (ice), and this error may be doubled by exposure to lower humidities at lower temperatures.

(iii) *Response of gold-beater's skin at atmospheric pressure*

A large part of this section deals with the lag of the G.B.S. In describing this response, use is often made of the conception of the half-change time τ, which is the time required to complete the first half of the change in indicated humidity after a sudden change of the relative humidity of the surrounding air. The reason for using the half-change time is that the lag curve cannot be represented by a simple exponential function like many other reactions. Under these circumstances the "half-change time" gives a much clearer perception of the response than the more precise mathematical expressions using a "response constant" k of a well defined function. The use of k will therefore be restricted to the parts dealing with the theoretical interpretation of the lag curve.

The response has been measured for changes at medium humidities at temperatures between + 18 and − 68° c. and at air velocities ranging from 1·2 to 34 metres per second. In figure 4 the logarithms of the half-change times are plotted against the reciprocals of absolute temperatures (T), resulting in parallel lines for the various air velocities. It is to be noted that the temperature coefficient of the response constant k (inversely proportional to τ) is only slightly smaller than that of the saturation pressure of liquid water.

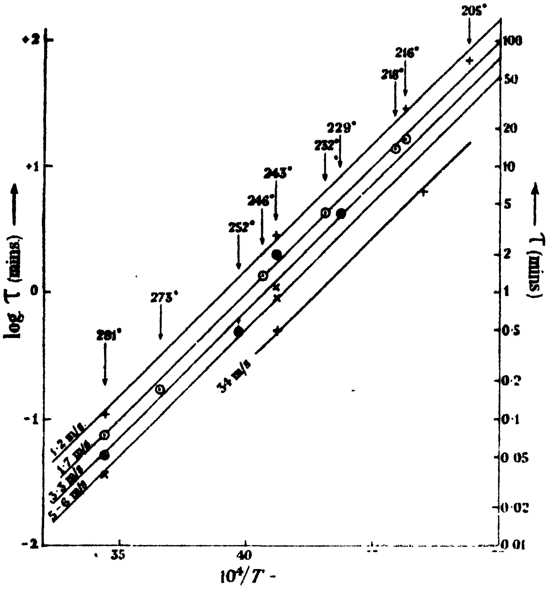


Figure 4. Response of gold-beater's skin at low temperatures and various air velocities (τ = half-change time).

Table 3 shows the half-change time at − 30° c. for different air velocities (u).

Table 3

u (m./sec.)	τ (min.)	$\frac{-d \log \tau}{d \log u}$
1·2	2·5	
		1·2
1·7	16·5	
		0·42
3·3	1·25	
		0·38
6	1·00	
		0·40
34	0·5	

It can be seen that, apart from the lowest value of u, τ is inversely proportional to $u^{0·4}$, which may be taken as an indication that the lag is almost entirely due to the diffusion of water molecules through the boundary layer surrounding the G.B.S. This is also borne out by the quantitative agreement between the half-change times found experimentally and those theoretically calculated on the basis that the lag is due to the boundary layer.

§4. CALCULATION OF THE LAG CURVE

Let dA/dh be the change of weight with relative humidity (h) per cm^2 of film or skin ($h=1$ at water saturation), so that $dA/dh=a\,.\,A_0$, where A_0 is the weight of film in gr./cm^2 and a may have different values in different regions of relative humidity.

$D=\frac{250}{P}\cdot\left(\frac{T}{273}\right)^{1\cdot 75}$ = diffusion constant of water vapour in air at a pressure of P millibars.

C = saturation concentration over water in gm./c.c.

δ = thickness of boundary layer.

n = number of sides of the film exposed to air (1 or 2).

As the change of weight of the film is due to the diffusion of water vapour through the boundary layer,

$$\frac{dA}{dt}=\frac{n\,.\,D\,.\,C}{\delta}(h_1-h)$$

at the time t after a sudden alteration of the humidity from h_0 to h_1. The change of indicated humidity (h) will then be

$$\frac{dh}{dt}=\frac{n\,.\,D\,.\,C}{\delta\,.\,a\,.\,A_0}(h_1-h)$$

or, integrated for changes of humidity not large enough to affect the value of a,

$$\frac{h_1-h}{h_1-h_0}=\exp.-\frac{n\,.\,D\,.\,C}{\delta\,.\,a\,.\,A_0}\,.\,t. \qquad \ldots\ldots(1)$$

Unfortunately, the simplicity of the function (1) is affected by δ not having a constant value along the length of the film, if the latter is in a position parallel to the air flow. The only case permitting of an exact evaluation is the case of a laminar air flow along the exposed film, when

$$\delta=4\cdot 5\sqrt{\frac{\nu\,.\,x}{u}}\text{ cm.},$$

where ν = kinematic viscosity, x = distance upwind from edge of film in cm., and u = air velocity in cm./sec.

Under these conditions the indicated change of relative humidity (h) is given by

$$\frac{h_1-h}{h_1-h_0}=\frac{1}{L}\int_0^L e^{-kt/\sqrt{x}}\,.\,dx, \quad \text{where} \quad k=\frac{n\,.\,D\,.\,C\,.\,\sqrt{u}}{4\cdot 5\,.\,a\,.\,A_0\,.\,\sqrt{\nu}}, \quad \ldots\ldots(2)$$

and where L is a length of the film or skin under consideration. The value of the integral (2) can be evaluated by means of Soldner's function (see Edwards, *Integral Calculus*, p. 334)

$$\frac{1}{L}\int_0^L e^{-kt/\sqrt{x}}\,.\,dx=\left(\frac{kt}{\sqrt{L}}\right)^2.\,\mathrm{li}\,(e^{-kt/\sqrt{L}})-\left(\frac{kt}{\sqrt{L}}-1\right).\,e^{-kt/\sqrt{L}}$$

in which li is the logarithmic integral. It is shown in curve B of figure 5 as a function of $kt/\sqrt{L}$, so that for a given value of $k/\sqrt{L}$ the time required for a certain change, e.g. 50% or 90%, can be directly obtained from this curve. It can be seen that curve B differs considerably from the functions exp. $-2kt/\sqrt{L}$ (curve A) and that it gives expression to the fact that the forward parts of the film or skin reach equilibrium earlier than those further down stream.

In applying equation (2) to the experimental data of the G.B.S. hygrometer used for the tests, the following numerical values must be employed:

$$n=2, \quad L=3\text{ cm.}, \quad A_0=1{\cdot}35\,.\,10^{-3}\text{ gm./cm}^2,$$
$$a=0{\cdot}2 \text{ (near } h=50\% \text{ R.H.)}.$$

The other physical data are given in table 4.

Table 4

T (°c.)	$10^6.C$ (gm./c.c.)	D	ν	u (cm./sec.)	$\frac{10^3k}{\sqrt{L}}$ (sec.$^{-1}$)	τ calc. (min.)	τ obs. (min.)	$\frac{\tau \text{ calc.}}{\tau \text{ obs.}}$
273	4·84	0·25	0·19	200	37	0·18	~ 0·2	~1
230	0·135	0·18	0·10	170	1·0	6·6	5	1·3
				330	1·4	5·0	3·6	1·4
				3400	4·3	1·5	1·3	1·2
216	0·031	0·16	0·09	120	0·18	37	29	1·3
214	0·026			3400	0·79	8·5	6	1·4

For the purpose of comparing the calculated lag values with those found experimentally and plotted in figure 4, it will again be an advantage to compare the times of the half-change τ. Not only can this point be determined experimentally with greater accuracy than points nearer the completed equilibrium, but, as will be seen shortly, there are difficulties in reducing the lag curves obtained experimentally to fundamental constants like $k/\sqrt{L}$. As shown in table 4, the calculated τ is larger than the observed one by a factor of 1·3 on the average, but the two τs are not strictly comparable.

On account of the limited amount of air at our disposal (less than 10 c. ft./min.), the experimental investigation was carried out in a very narrow channel of X-section 0·4 × 2·5 cm², and the flow was thus not laminar, but highly turbulent at the entrance of the channel.

Tables 5 *a* and 5 *b* show typical lag curves at a temperature of $-29°$ c. taken at an air velocity of 4·2 m./sec. The humidity range chosen (from 90% to 57% R.H. with respect to ice) corresponds to medium humidities with respect to water, where the value of a is very constant. Comparison of the third columns in both tables show that the G.B.S.—unlike the hair hygrometer—follows the same response function independent of the direction of the humidity change.

It can be seen from tables 5 *a* and 5 *b* that the lag curves obtained experimentally are neither simple exponential like curve A, nor do they follow curve B, which is based on the assumption of laminar flow (equation (2)). They correspond to a curve of type C (figure 5). Curve C is based on the assumption that the thickness δ of the boundary layer is proportional not to $\sqrt{x}$ but to x, resulting in the equation

$$\frac{h_1-h}{h_1-h_0}=\frac{1}{L}\int_0^{x=L} e^{-\frac{kt}{bx}}\,.\,dx, \quad \text{where } k=\frac{n\,.\,D\,.\,C\,.\,\sqrt{u}}{4{\cdot}5\,.\,a\,.\,A_0\,.\,\sqrt{\nu}}, \quad \ldots\ldots(3)$$

and where the constant b may have values up to $1/\sqrt{L}$, which means that δ cannot exceed the value it would have under laminar flow conditions. Curve C has been constructed for $b=1/\sqrt{L}$. Equation (3) can be integrated by using again Soldner's "logarithmic integral",

$$\frac{1}{L}\int_0^L e^{-\frac{kt}{bx}}\,.\,dx=e^{-\frac{kt}{bL}}+\frac{kt}{bL}\,.\,\mathrm{li}\left(e^{-\frac{kt}{bL}}\right).$$

Table 5 *a*. $T=-29{\cdot}0^\circ$ c.

t (min.)	h observed	$\frac{h_1-h}{h_1-h_0}$	$k/\sqrt{L}$ according to "A"	$k/\sqrt{L}$ according to "B"	$k/\sqrt{L}$ according to "C"
0	$90\%=h_0$	1·000			
0·25	83·4	0·798	0·48	0·48	0·28
0·5	78·9	0·660	0·41	0·45	0·29
0·75	75·5	0·557	0·39	0·44	0·30
1	72·8	0·472	0·38	0·44	0·30
1·5	69·0	0·356	0·35	0·43	0·305
2	66·3	0·274	0·33	0·41	0·30
3	63·1	0·175	0·29	0·38	0·29
4	61·5	0·125	0·26	0·34	0·29
5	60·4	0·090	0·24	0·32	0·27
6	59·0	0·047	0·25	0·34	0·29
10	$57{\cdot}5\%=h_1$				

Table 5 *b*. $T=-28{\cdot}5^\circ$ c.

t (min.)	h observed	$\frac{h_1-h}{h_1-h_0}$	$k/\sqrt{L}$ according to "A"	$k/\sqrt{L}$ according to "B"	$k/\sqrt{L}$ according to "C"
0	$57\%=h_0$	1·000			
0·25	65·4	0·770	0·54	0·58	0·32
0·5	69·2	0·614	0·46	0·54	0·36
0·75	71·5	0·555	0·39	0·45	0·30
1	74·2	0·475	0·37	0·44	0·30
1·5	78·4	0·349	0·36	0·43	0·31
2·25	82·2	0·232	0·33	0·41	0·31
3	84·6	0·163	0·30	0·40	0·31
4	86·2	0·109	0·28	0·37	0·32
5	87·6	0·070	0·27	0·36	0·30
9	$90\%=h_1$				

As the actual lag curves agree in form with the curve from equation (3) (see column 6, tables 5 *a* and *b*), it must be concluded that, under the experimental conditions, the boundary layer along the G.B.S. is affected by the initial turbulence and has a more or less linear form. This would also explain the results of table 4, which show a smaller value for the experimentally found half-change time than that obtained under the assumption of laminar flow.

§ 5. THE RESPONSE OF GOLD-BEATER'S SKIN AT DIFFERENT RELATIVE HUMIDITIES

Figure 6 shows the response constant $k/\sqrt{L}$ ($=0{\cdot}28/\tau$) for small changes at various humidities at 18° c. The curve exhibits a marked maximum between 50 and 65% R.H. This behaviour is quite different from that of the hair hygrometer, for which the half-change times decrease throughout towards higher humidities. Figure 7 shows that this phenomenon is substantially the same at lower temperatures (−21° c. and −30° c.). The maximum for −21° c. lies again in the same region, if relative humidity is taken with respect to supercooled water; at −30° c. the falling branch is unobtainable on account of ice saturation.

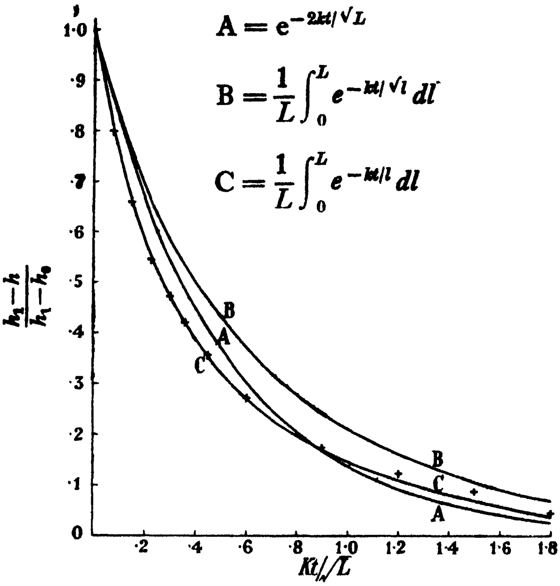


Figure 5. Calculated reponse curves of gold-beater's skin and observed response (+).

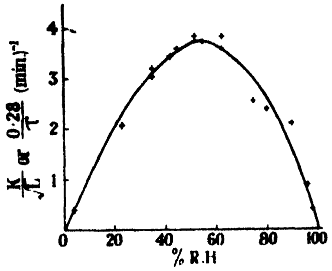


Figure 6. Response in different regions of relative humidity at 0° c.

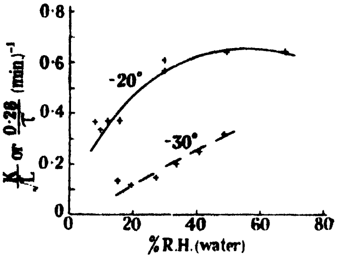


Figure 7. Response in different regions of relative humidity at −20° c. and at −30° c.

§ 6. CONCLUSIONS CONCERNING GOLD-BEATER'S SKIN

While gold-beater's skin is much faster in response than hair, it is still slow at temperatures below −40° c., as can be seen from the data of figure 4, which give the half-change times at medium relative humidities where the response is a maximum. At low humidities it is much slower still. It would even appear from figures 6 and 7 that the response ceases at relative humidities approaching zero. Observations both in the laboratory and in the stratosphere (by Mr. A. W. Brewer) confirm that gold-beater's skin loses all response after prolonged exposure to zero humidity and that it only regains it after exposure to high humidities above −15° c.

Whether a reduction of the atmospheric pressure would bring an improvement of the normal response could not be ascertained experimentally. Rough calculations under the assumption that the internal diffusion coefficient of gold-beater's skin is similar to that of hair would lead to the conclusion that, while reduction of pressure would improve the response at medium to high humidities, it would not have a great effect below 20% R.H., where a reduction of pressure would make internal diffusion the dominant factor.

The hysteresis effect after exposure to humidities below 20% R.H. is less than in the case of hair. However, it is still far from negligible and is likely to affect the accuracy of the readings.

These results show that, while gold-beater's skin is an improvement on hair, its limitations at low humidities are considerable.

§7. THE ELECTROLYTIC HYGROMETER

Tests similar to those with gold-beater's skin were made with units of the electrolytic hygrometer of Dunmore (1939), which is based on the resistance of a thin electrolyte film containing lithium chloride. These tests appeared to be all the more necessary as the instrument was apparently used at low temperatures without any calibration except at ice saturation, other humidities being extrapolated from the behaviour at room temperature.

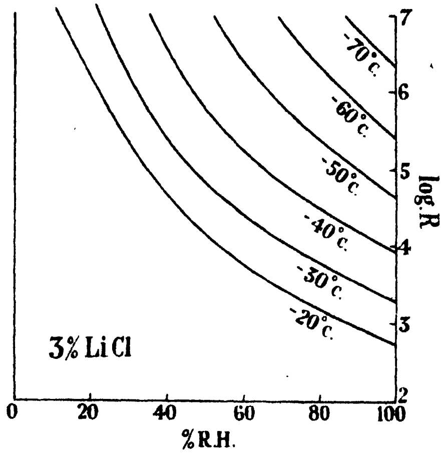


Figure 8. Resistance (R) of Dunmore unit with 3 % LiCl at different temperatures and humidities.

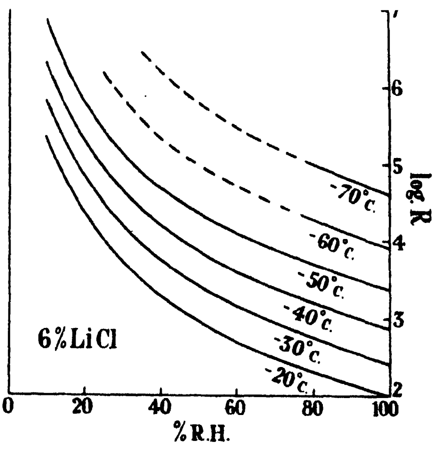


Figure 9. Resistance (R) of Dunmore unit with 6% LiCl at different temperatures and humidities.

The resistance measurements at low temperatures and at different humidities, which were carried out with units obtained from U.S.A., showed immediately that this procedure was not admissible. The temperature coefficient of the resistance varies enormously at different relative humidities, as can be seen from figure 8, for a unit made from a 3% LiCl solution. (For composition of solutions see Dunmore (1939), p. 705). For a change from −20° c. to −40° c. the resistance of these units changes

at ice saturation by a factor of 14,
at 70% R.H. (ice) by a factor of 25,
at 40% R.H. (ice) by a factor of 70.

Such variations make an extrapolation from room-temperature conditions impossible, and these differences are greater still for units containing smaller percentages of LiCl, such as are used in the composite hygrometer described by Dunmore (1939), figures 2 and 3, and Diamond, Hinman and Dunmore (1940).

On the other hand, units made from solutions with more than 6% LiCl did not show this variation with humidity of the temperature coefficient of the resistance (see figure 9). This fact, as will be discussed later, makes possible the construction of temperature-independent hygrometers of this type.

The electrolytic hygrometer does not show any hysteresis after exposure to low humidities. This great advantage over hair and gold-beater's skin is probably due to the amorphous structure of the film. This, on expansion or contraction, does not set up the internal stresses responsible for a hysteresis effect (see **Barkas,** 1942) which would occur in semi-crystalline structures like cellulose, hair and gold-beater's skin.

Response of the Dunmore hygrometer

The response of the Dunmore units which, on account of the form of the humidity chamber, had to be tested with the air flow parallel to their axes, is shown in table 6 for a unit of 20 wire turns made from 3% LiCl solution at a temperature of −28° c.

Table 6

Air velocity (m./sec.)	Half-change times for changes from : 57% to 90% R.H. (min.)	90% to 57% (ice) (min.)
0·4	4·5	12
2·7	2·0	5
5·2	1·1	3
6·7	0·7	-
44	0·2	0·5

It can be seen that the change from low to high humidity is about 2·5 times faster than that in the opposite direction, and that the response depends on the air velocity (u) approximately as $u^{0\cdot65}$. The dependence on the air velocity indicates that the lag is due to the diffusion of water vapour through the boundary layer. The reason for the influence of the direction of the humidity change can also be easily understood. The various parts of the electrolytic hygrometer unit may be considered as "resistances in parallel", so that the total resistance is substantially determined by the parts of lowest resistance, i.e. of highest moisture content. During the change from low to high humidity, the total resistance is thus at first governed by the most exposed parts, while during the opposite change the resistance depends largely on the least ventilated parts of the unit, and consequently shows less response during the initial changes.

The variation of the response with temperature is given in table 7 for changes from low to high humidity at an air velocity of 44 m./sec. (unit: 20 turns, 6% LiCl).

Allowing for differences in flow velocity, temperature and arrangement, these times are in line with the figure given by Diamond, Hinman, Dunmore and Lapham, who find a half-change of 0·12 minute at 5 m./sec. at 0° c.

As these half-change times are increased by a factor of 6 for ventilation speeds of 5 m./sec. (normal for radio-sonde work) and by another factor of 2·5 for changes from high to low humidity, it is apparent that the lag of this hygrometer, when

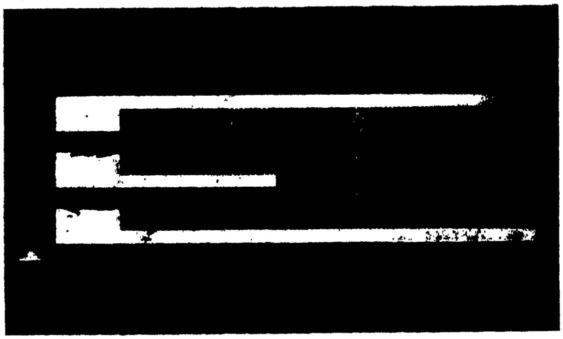


Figure 15. Electrolytic hygrometer element (for temperature compensation, made by Scientific Instrument Research Association).

Figure 10. Electrolytic hygrometer with temperature compensation and wide range of humidities (circuit, see figure 11).

used in the radio-sonde, cannot be neglected at temperatures below −40° c. It may, however, be expected that in this case considerable improvements may be caused by the low density of air in the upper atmosphere.

Table 7

Temperature (°c.)	Half-change (min.)
−20	<0·1
−30	0·2
−40	0·7
−50	2
−60	appr. 6
−70	appr. 25

§8. MODIFICATIONS OF THE ELECTROLYTIC HYGROMETER

Improvements were directed towards designing a temperature-independent electrolytic hygrometer (thus eliminating the necessity to calibrate every instrument at low temperatures) and towards increasing the response.

The first aim was achieved by compensating the temperature effect by means of a similar electrolytic resistance kept at ice saturation. This could be done for units made from 6% LiCl solution, where, as stated before, the temperature effect is the same at all humidities.

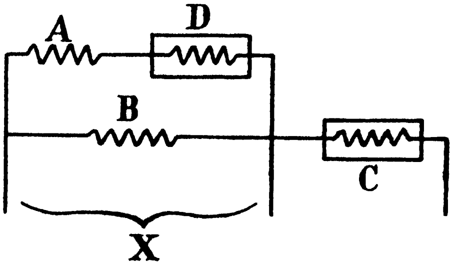


Figure 11. Circuit of electrolytic hygrometer with temperature compensation.

To render the scale readings more linear, the compensated hygrometer, shown in figure 10, was composed of four resistances, all made from polyvinylacetate with 6% LiCl, as follows:

Resistance A = 20 turns at 16 turns per inch,
B = 2 turns at 16 turns per inch,
C = 1 ring, 1/16 inch wide,
D = 1 ring, 1/8 inch wide.

These arrangements had the result that A : B : C : D = 1 : 7 : 42 : 21 if all were kept at the same humidity. In the hygrometer, the circuit of which is shown in figure 11, A and B were exposed to the air, while C and D were kept in an atmosphere saturated with respect to ice. The saturation was produced by keeping C and D in an annular closed space containing a ring of filter-paper soaked with saturated K_2SO_4 solution. This produced ice saturation below −5° c., and at the same time prevented the "flooding" of the LiCl film at room temperature which would occur with pure water.

The resistance ratio X/C was found to be independent of the temperature between $+5°$ and $-60°$ c. where X and C are the resistances of the unit and the compensator respectively. Figure 12 shows a series of calibrations, carried out over a fortnight, with the compensated hygrometer shown in figure 10. In this test the relative humidities (with respect to ice) were measured with an early model of the dew-point hygrometer by Dobson, Brewer and Cwilong, kindly lent by the Meteorological Office of the Air Ministry. It is difficult to say whether the scatter of the calibration points was due to the electrolytic hygrometer or to the difficulty in operating this dew-point instrument, which was not as sensitive as later models. But apart from this, the constancy of the resistance ratio was well established, thus providing the basis for electrolytic hygrometers with temperature-independent calibration.

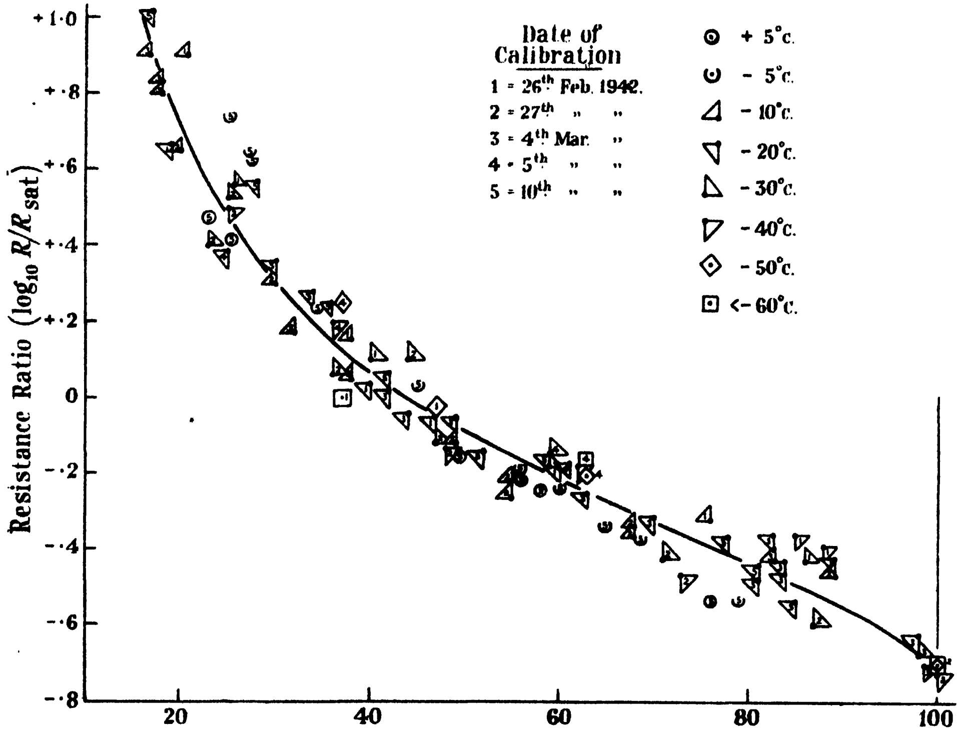


Figure 12. Calibration curve of compensated electrolytic hygrometer.

To increase the response of the electrolytic units, the electrodes were platinized directly on to the glass holders, as shown in figure 13. This unit was produced by painting the electrode pattern with platinizing solution on a glass bulb with subsequent heating. These units had the advantage that much less hygroscopic material adheres to the smooth surface than was retained by the wire-covered Dunmore units. Consequently, less moisture was required to produce equilibrium, which resulted in a faster response. Figure 14 shows the response curves for this type of hygrometer unit at a temperature of −50° c. at an air velocity of 18 m./sec. The half-change times were 0·4 and 0·7 minute, according to the direction of the humidity change, an average improvement by a factor of 10 over the performance of the wire-wound units.

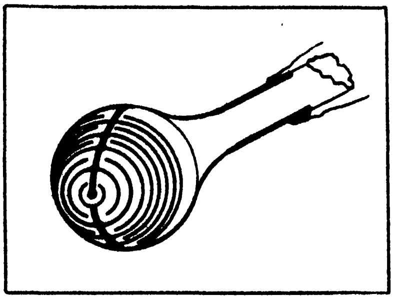


Figure 13. Electrolytic hygrometer element (platinum on glass bulb)

Still faster response was obtained with platinum-on-glass electrodes, kindly made for me by the British Scientific Instrument Research Association, which are shown in figure 15. The close spacing of the electrodes and the comparatively large electrode area made it possible to use films of less than 0·001 mm. thickness. These films gave manageable resistances, and half-change times of about 0·5 minute were obtained at −55° c. at an air velocity of 4 m./sec.

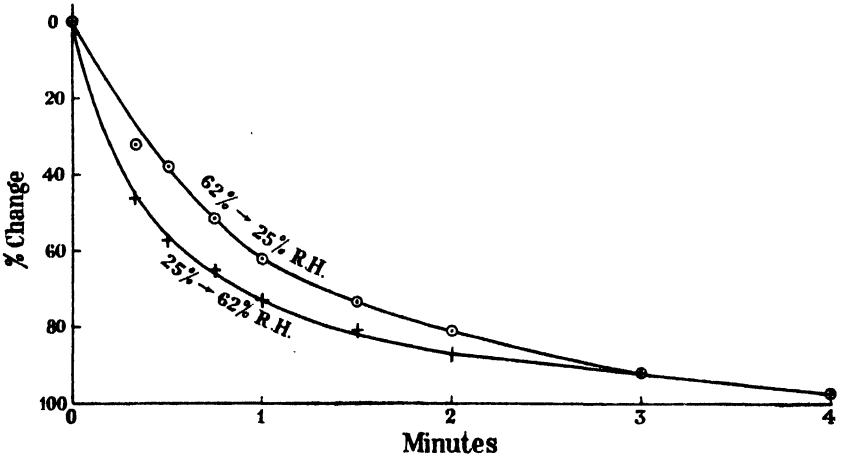


Figure 14. Response curve at −50° c. of electrolytic hygrometer unit shown in figure 13 (air ventilation 18 m./sec.).

At this point progress came to a standstill through technical difficulties the cause of which was not recognized until very much later. It was found that the platinum-on-glass units showed a slow but continuing change in calibration with time, which affected not only the absolute value, but also the temperature coefficient of the resistance. Systematic investigations revealed that the films behaved as if LiCl was constantly removed from the electrolyte.

Nor was this effect of a simple nature. In films with high LiCl content (6%), the process seemed to be accompanied by the formation of microscopic needles in the film. These formed only if the film was exposed to high humidities, but their appearance could be indefinitely postponed by keeping the units in a desiccator. No indication of their chemical composition was found, but similar needles were also obtained for mixtures of LiCl with other binders such as gelatine, gum arabic, and other types of polyvinylacetates when used on glass, while these needles could not be detected on the Dunmore units from U.S.A. Using coating solutions of less than 3% LiCl, no needles were found, but the apparent disappearance of LiCl was observable also with these very thin films. The cause of this seemed to elude all efforts at detection, until a specimen which had been kept for more than a year showed under the miscrocope that small cubic crystals had been formed. It then became apparent that at least one of the disturbing processes must have been the ion exchange between the lithium of the film against the sodium of the glass, the cubic crystals being most likely NaCl.

It is quite possible that, if this work should again be taken up, a technical solution may be found, e.g. by using quartz glass or fused alumina as bases for the platinum patterns. If this should result in stable films, the way would be open to constructing a very simple hygrometer of sufficiently fast response at all temperatures down to – 70° c. free from hysteresis effects at low humidities, which would not require low-temperature calibration, and which could easily be adapted for use both on aircraft and radio-sonde.

§ 9. AN OPTICAL HYGROMETER

During attempts to throw some light on the difficulties incurred with the very thinnest films used for the electrolytic hygrometer, it was observed that these films gave interference colours in reflected light and that these colours changed markedly under varying conditions of relative humidity. The effect seemed to be sufficiently characteristic to serve as the basis for a new type of hygrometer.

In order to test the behaviour of hygroscopic films at low temperature, an arrangement was used which is shown in principle in figure 16. A coated lead-glass disc of 10 mm. diameter was held at an angle in a metal frame within the central tube of the calibration apparatus shown in figure 1. Light from a 12-v. lamp (L) was condensed into a narrow beam by means of a microscope eye-piece (E) which was filled with a mixture of glycerine and water so as to absorb heat radiation. The "interfered" light was reflected by the hygroscopic surface (A) into a photocell (P) of the caesium type. A thermocouple was fixed to the back of the lead-glass disc (A) to observe any differences of temperature between the bath and the hygrometer unit, caused by absorption of light at the blackened back of the lead-glass disc.

The intensity of the light source was adjusted so as to give a definite photocell current, when the light was reflected by the back of the spring-loaded shutter (St) which, when released, would cover the hygroscopic surface and thus act as a standard reflector. By removing the box with the photocell it was possible to view directly the changes of colour of the reflected light.

Hygroscopic films of very uniform thickness were produced by spinning the solutions on discs of lead-glass of high refractive index. Extremely vivid colours were obtained, the variability depending on the LiCl content. By adjusting the concentration and viscosity of the solutions, and by varying the speed of the

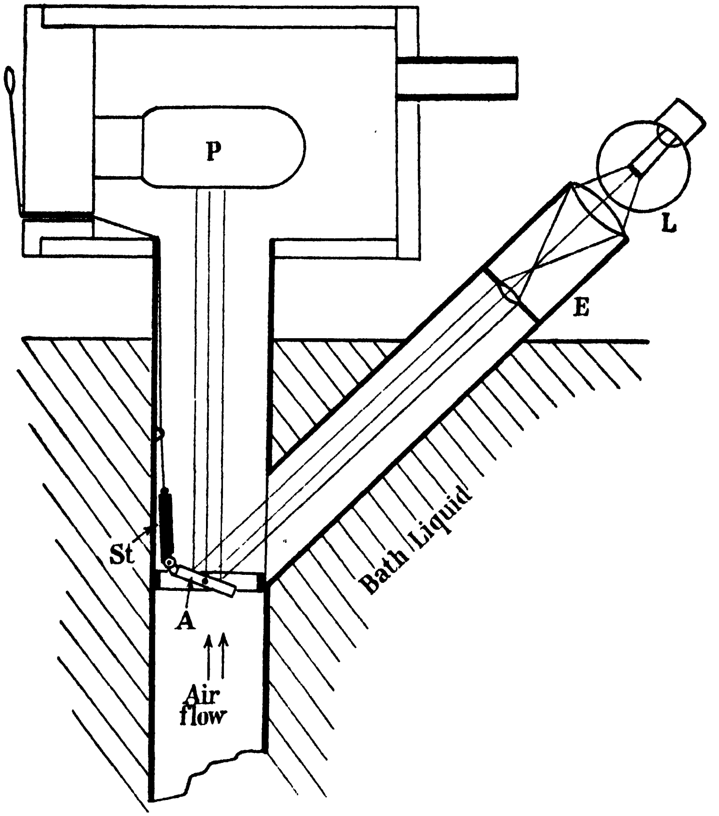


Figure 16. Arrangement used in combination with testing apparatus (see figure 1) for calibration and response measurements of " optical hygrometer " elements. L, lamp ; E, lens ; A, hygroscopic film surface ; St, standard reflector ; P, photocell.

rotating discs, any desired initial thickness of film could be produced. Examples of interference colours obtained for Solvar films on lead-glass, and their variation with relative humidity, are shown in table 8.

Table 8

% R.H.	6 % LiCl solution	1 % LiCl solution
0	Yellow	White
10	Orange	
20	Purple	
30	Blue	Yellow
40	Blue-green	
50	Yellow-green	Orange
60	Orange	
70	Red	Red
80	Blue	Purple
90	Yellow	Blue
96	Red	Green

A more quantitative investigation of the swelling properties was based on the intensity changes of monochromatic light reflected from the hygroscopic film on lead-glass. It was found that at any given relative humidity with respect to liquid or supercooled water, the thickness of the films containing LiCl remains unaltered within the limits of error between the temperatures of +15 and −50° c. This fact is of great importance, because it forms the basis for a hygrometer with temperature-independent calibration.

Table 9 gives the changes of thickness with relative humidity found for films of different LiCl contents (which are here given in weight % of dry film). The thickness at 0% relative humidity is defined as unity. The region between 0 and 35% R.H. shows an approximately linear response.

Table 9

% R.H.	Weight % LiCl in dry film					
	0	8·4	15·5	21·6	27·0	35·5
0	1·00	1·00	1·00	1·00	1·00	1·00
35	1·01	1·025	1·05	1·12	1·16	1·24
52	1·02	1·06	1·10	1·19	1·28	1·32
76	1·06	1·15	1·24	1·43	1·52	1·55
96	1·17	1·33	1·78	2·15	2·30	2·56

These observations were utilized for the development of a temperature-independent optical hygrometer (see figure 17) based on the change of intensity of reflected monochromatic light.

Parallel light from a 6-v. 6-w. bulb (L) was reflected by a coated lead-glass disc (A) of ½ inch diameter to a vacuum photocell of the caesium type. Caesium photocells are particularly sensitive to red light, so that the photocell even without a colour filter indicated essentially the intensity change of red light. The monochromatism was further improved by a filter of methyl violet in front of the photocell. Light from the same source fell through an adjustable slit (S) on a second photocell, which served to compensate against changes in the light intensity of the lamp. Both photocell currents were suitably amplified, using a cathode follower arrangement (Sowerby, 1944). A method (Glückauf, 1945) for measuring directly the ratio of the illuminations of the two photocells was not used because vacuum photocells proved to be more reliable than the otherwise needed gas-filled cells. Changes due to relative humidity were read with the millivoltmeter (M) of high resistance. (For satisfactory working, resistance in the voltmeter circuit should be made a good deal higher than the cathode follower resistance.) The width of the slit S is adjusted so that at zero humidity no current flows through the voltmeter.

The change of light intensity I of monochromatic light with the film thickness F(given in terms of the wave-length) should give a cosine wave, but in practice, with incompletely monochromatic light, the amplitudes of the cosine waves gradually fall off (see figure 18). In order to compensate for the very much higher expansion of the films per unit change of relative humidity at high humidities (see table 9) it is of advantage to choose the thickness and composition of the film in such a way that the zero-humidity thickness has a value where dI/dF is a maximum (e.g. $\lambda/8$, $3\lambda/8$, $5\lambda/8$, etc.) and that the 96% saturation thickness has values with minimum dI/dF (e.g. $\lambda/4$, $2\lambda/4$, $3\lambda/4$, etc.). Two useful ranges are indicated in figure 18: from $\lambda/8$ to $\lambda/4$ and from $3\lambda/8$ to $\lambda/2$. These require materials of a composition giving, between 0 and 96% R.H., an expansion of 100% and 33% respectively.

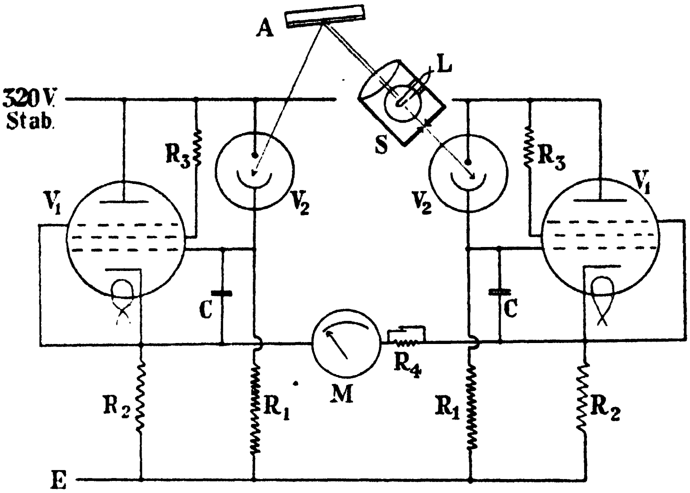


Figure 17. Arrangement and electrical circuit of optical hygrometer. A, hygroscopic surface; L, 6-watt lamp; S, shutter, adjustable. Electrical circuit: V_1, H.F. pentode; V_2, Baird vacuum photocell (Cs-type); $R_1 = 100$ M; $R_2 = 100{,}000$; $R_3 = 30{,}000$; R_4 according to resistance of meter M; $C = 0{\cdot}01$ μF.

According to table 9 this requires film materials containing 19% and 8·4% LiCl respectively in the dry film. As both films require, independently of their zero-humidity thickness, the same amount of moisture and, consequently, have the same response characteristics, there is no advantage in choosing the thinner film. As the hardness of the films increases with decreasing LiCl content, it is actually better to use the thicker film. A zero thickness of $3\lambda/8$ was therefore used throughout.

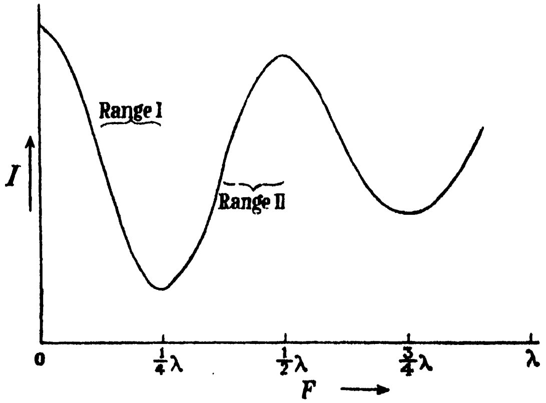


Figure 18. Variation of light intensity (red) I reflected from hygroscopic film. Abscissa: thickness of film in (average) wave-lengths of light.

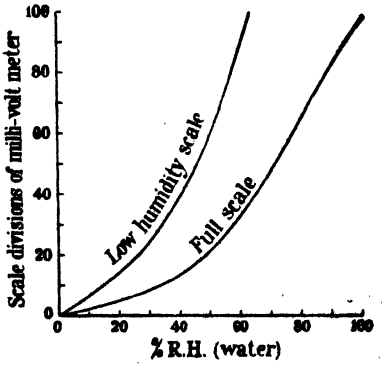


Figure 19. Calibration curves of optical hygrometer (see figure 17).

By using a range from $5\lambda/8$ to $3\lambda/4$ it would have been possible to dispense with LiCl altogether and to use pure Solvar films. But not only were the amplitudes of the *I–F* curve much smaller in this region, which would have reduced the proportional change in light intensity, but there were also other difficulties. It was found that the Solvar films without any LiCl showed a slight variation of their calibration curve with temperature which, in view of the wide temperature range for which the instrument was intended, could not be neglected. (Physico-chemically this phenomenon would mean that the heat of swelling of the pure Solvar films is not negligible.) For a more moderate range of temperatures, down to about −30° c., this temperature effect is almost negligible. For this purpose, the use of pure Solvar films offers considerable advantages as regards stability, for films containing LiCl are destroyed by contact with liquid water.

In order to give greater scale uniformity in different regions of humidity, two different voltmeter sensitivities were used. This is particularly useful for work at very low temperatures, where the high relative humidities with respect to water do not occur on account of ice saturation.

An instrument built on these principles with the circuit of figure 17 gave calibration curves as shown in figure 19. It can be seen that the arrangement by which the strongly expanding high-humidity part coincides with the least sensitive part of the interference curve does indeed lead to a fairly linear curve in the high-humidity region.

It was at first feared that ambiguities in the indication might arise for air humidities higher than 96% relative humidity, when the light intensity, after having reached the maximum, does again decrease. However, in practice this was not found to be the case. The radiation from the lamp heats the reflecting film to a temperature slightly above that of the surrounding air, so that under constant conditions of ventilation the relative humidity at the hygroscopic surface is always a small and constant fraction below the actual air humidity, and relative humidities above 96% do not occur at the hygroscopic surface.

At room temperature and with moderate ventilation the response of the hygrometric film is instantaneous. This is obvious, considering that the amount of moisture required for a change, e.g. from 52 to 76% T.H., is only $3\,\mu$ gr./cm^2, and from 35 to 52% R.H. is only $1\,\mu$ gr./cm^2, so that exchange with less than 1 cm. of air layer is required to reach equilibrium.

Experiments at −50° c. at atmospheric pressure, in which a ventilation of 10 m./sec. was directed against the back of the lead-glass disc (see arrangement in figure 16), gave half-change times of about 20 seconds. At high altitudes and low air densities a further increase in response may be expected.

The only technical difficulty which has so far prevented the use of the optical hygrometer for practical purposes is the lack of a suitable base material of high refractive index. When covered with a hygroscopic film, lead-glass becomes slowly tarnished, a process which proceeds faster the higher the relative humidity of the surroundings. This has the result that the calibration curve is subjected to a slow continual shift, making necessary calibrations at daily intervals. During the war years it was not found possible to obtain glass of high refractive index made from materials other than lead, nor was fused alumina of sufficiently high purity obtainable.

In the meantime, the development of a low-temperature dew-point hygrometer by Dobson, Brewer and Cwilong, with its obvious advantage of not requiring calibration, has somewhat reduced the interest in absorption hygrometers for high-altitude hygrometry. In particular, the very low relative humidities which are found above the tropopause (Glückauf, 1945 a; Dobson *et al.*, 1946) are far more reliably measured with a dew-point instrument than with absorption hygrometers which, at these very low humidities, are either too slow in response (goldbeater's skin), unsuitable on account of a too high resistance (electrolytic hygrometer), or comparatively less sensitive (optical hygrometer). It is likely, however, that, with non-tarnishing base materials becoming available, the optical hygrometer, on account of its very fast response, its simple construction, its independence of temperature and its suitability as a direct-reading or recording instrument, may have some future.

§ 10. ACKNOWLEDGMENTS

My sincere thanks are due to Sir Nelson K. Johnson, Director of the Meteorological Office, Air Ministry, and to Dr. G. M. B. Dobson, F.R.S., for their great interest in this work; to the Gassiot Committee of the Royal Society for the appointment to the Mackinnon Research Studentship, 1942–1944; to Mr. A. J. Philpot, O.B.E., M.A., Director of the British Scientific Instrument Research Association for advice and technical help; to the Durham Colleges in the University of Durham, and especially to Professor F. A. Paneth, for their interest and for generous laboratory facilities.

REFERENCES

BARKAS, W. W., 1942. *Trans. Faraday Soc.*, **38**, 194.

DIAMOND, H., HINMAN, W. S., Jr. and DUNMORE, F. W., 1940. *J. Res. N.B.S., Wash.*, **25**, 327.

DOBSON, G. M. B., BREWER, A. W. and CWILONG, B. M., 1946. *Proc. Roy. Soc.*, A, **185**, 148.

DUNMORE, F. W., 1939. *J. Res. N.B.S., Wash.*, **23**, 701.

FINDEISEN, W., 1937. *Wiss. Abh. Reichsanst. für Wetterdienst*, **2**, no. 11, part 2.

GLÜCKAUF, E., 1944. *Quart. J. Roy. Met. Soc.*, **70**, 293 ; 1945 a. *Ibid.*, **71**, 110 ; 1945 b. *J. Sci. Instrum.*, **22**, 34.

SOWERBY, J. McG., 1944. *J. Sci. Instrum.*, **21**, 42.

A THEORY OF FLICKER NOISE IN VALVES AND IMPURITY SEMI-CONDUCTORS

By G. G. MACFARLANE,
Telecommunications Research Establishment, Ministry of Supply, Malvern

MS. received 16 *October* 1946 ; *read* 21 *February* 1947

ABSTRACT. A theory of contact noise at low frequencies is described. It is assumed that this noise is due to diffusion of clusters of mobile impurity centres on to the contact surface, as in Schottky's theory of flicker effect in valves. A cluster disappears in time due to ionization and consequent ionic conduction away from the contact surface. This diffusion-conduction process is used to derive a formula for the flicker noise which is applicable to emission from oxide-coated filaments of valves and to contacts between particles of impurity semi-conductors, as in lead sulphide photo-conductive cells and rectifiers. The spectral power density of the noise is found to depend on current j and frequency f as j^{x+1}/f^x, where $1<x<2$. Experimental results for PbS cells, oxide emitters and copper-oxide rectifiers are found to be in good agreement with this law. Carbon resistors are also found to obey this law, and it is suggested that the same mechanism of diffusion of clusters of impurity atoms on to the contact surface between crystals is responsible for the effect.

§1. INTRODUCTION

Recent experimental studies of the power spectra of low-frequency noise observed in lead-sulphide photo-conductive cells, carbon resistors and copper oxide rectifiers, described by Harris, Abson and Roberts (1947), have shown that when current is flowing the power density of the noise increases rapidly as the frequency is reduced. A similar phenomenon is well known for valves with oxide-coated filaments (Johnson, 1925; Schottky, 1926) and is called *flicker effect.* This frequency-dependent noise has also been observed in silicon crystal rectifiers (Miller *et al.*, 1946) and in carbon granule microphones (Christensen and Pearson, 1936). In each case this extra noise, which we shall call flicker noise, is found to obey a law of the form

$$\overline{\Delta j^2} \propto \frac{j^m}{f^n},$$

where $\overline{\Delta j^2}$ is the mean square fluctuation of current, j the mean current and f the frequency; n ranges from 0·6 to less than 2 and m from 1·5 to 3.

Flicker noise is not the only species of noise at low frequencies. Thermal agitation noise (Johnson noise), which is independent of current, and shot noise, which is proportional to current, may also be present. In both these cases the spectral density is constant, independent of frequency. At low frequencies ($<10^3$ to 10^4 c.p.s.) flicker noise may be tens or even hundreds of times greater than Johnson or shot noise. At higher frequencies, flicker noise becomes negligible.

Consideration of the known systems in which flicker noise occurs reveals the common feature of contacts between different particles or different materials.

In most cases one of the materials at a contact is a semi-conductor. Examples are (i) the valve, in which the contact is between an excess semi-conductor, viz. an oxide of a rare earth having excess metal, and vacuum; (ii) lead-sulphide cell, which consists of a great number of micro-crystals in contact, the lead sulphide having excess lead; (iii) the copper-oxide rectifier, which consists of a deficit semi-conductor in contact with a thin insulating layer separating it from metallic copper; (iv) the carbon resistor and the carbon microphone, which consist of a multitude of carbon granules in contact. Now flicker noise occurs only when current is flowing. This suggests that it is due to variations in the conducting properties of contacts.

A theory of flicker noise in valves was described by Schottky in 1926. He ascribed the effect to fluctuations in the surface layer of foreign atoms on the cathode. The elementary event is the coming and going of separate foreign atoms or molecules, and the time constant of the event is identified with the duration of the stay of an individual atom or molecule on the surface. Each foreign atom is assumed to reduce or increase the effective work function of the surface by the same amount. Following Langmuir, Schottky accounted for this change in effective work function by supposing that a foreign atom on the surface was polarized and therefore produced an electric double layer on the surface (Schottky and Rothe, 1928). As a basis for estimating the spectral density function, Schottky assumed that the life of an adatom is controlled by a diffusion process independent of the current flowing. The correlation function for the process, which is the probability distribution of life times of adatoms on the surface, is then a pure exponential, $\exp(-qt)$, where $1/q$ is the average sojourn of an adatom on the surface. This leads to the law

$$\overline{\Delta j^2} \propto \frac{j^2}{\omega^2+q^2} \simeq \frac{j^2}{\omega^2} \text{ for } \omega \gg q,$$

where $\omega = 2\pi f$. Schottky attempted to fit this law to observations of Johnson but the fit was not very good. The experimental law was more nearly

$$\overline{\Delta j^2} \propto \frac{j^2}{\omega^{1\cdot 25}}.$$

Recently R. L. Sproull (1945) has described a surface diffusion process which satisfactorily explains the initial rapid fall in emission observed when current is suddenly drawn from a valve. At the end of his article, Sproull suggested that this decay effect might be connected with flicker effect in valves. It is our purpose to explain the connection and to use the diffusion-conduction theory to derive an expression for flicker noise. It will be shown that the theory leads to a spectral-density law that is in good agreement with experiment. Moreover it is suggested that a similar mechanism is responsible for flicker noise in other systems in which there are contacts involving impurity semi-conductors such as lead-sulphide cells, rectifiers and non-metallic resistors.

It is, however, a salient feature of our theory that the observed flicker noise arises from the diffusion of clusters of impurity centres (atoms of barium in the case of a BaO emitter) on to the emitting surface. These clusters appear only at a relatively small number of points on the emitting surface. Only in this way does it seem possible to account for the deviation of the frequency law from $1/\omega^2$.

This is in contrast to Schottky's theory, in which the fluctuations are due to the random variations in the surface density of adatoms considered as free molecules of a gas.

§ 2. THE DIFFUSION-CONDUCTION THEORY

For the present we shall restrict our discussion to the case of emission from an oxide-coated cathode. Here we have a semi-conducting mass of, say, BaO emitting electrons into a vacuum. From electron-microscopic studies it is known that the emission of electrons does not take place uniformly over the surface of the cathode but in patches. This is most apparent in the process of activation when barium atoms erupt on to the surface at scattered regions (Ahearn and Becker, 1938).* The occurrence of these eruptions means that fluctuations in the density of barium atoms on the surface will be considerably greater than the square root of the average surface density, as assumed by Schottky. The correlation function for the fluctuations can then be appreciably different from the simple exponential, as we shall show.

Let us now consider the process by which adatoms of barium appear on, and then disappear from, the surface at a brightly emitting patch when an electric field is applied.

Barium atoms diffuse on to the surface when there is a concentration gradient at the surface tending to move Ba atoms out from the interior. Barium can be transported away from the surface either by evaporation, which we shall neglect, or by ionic conduction. While barium atoms remain on the surface they are assumed to be strongly polarized and, therefore, to form a dipolar layer over the emitting patch. The strength of the double layer is proportional to the surface density of adatoms provided this is less than the density required to produce maximum emission. The effective work function is reduced by the presence of the double layer by an amount which is directly proportional to the strength of the double layer. The rate at which ionized barium atoms leave the surface is taken to be proportional to the applied electric field and, therefore, to the electronic current. The diffusion rate is proportional to the number of barium atoms that have left the surface. The diffusion-conduction equation is, therefore,

$$\frac{dN}{dt} = -\frac{\alpha}{e} j + p(N_0 - N), \qquad \ldots\ldots(1)$$

where N is the number of adatoms per cm^2 of surface at time t, j is the thermionic current density, e is the electronic charge, α is the ratio of ionic to electronic conductivities, p is the probability that a Ba atom shall move from the surface to a distance h from the surface in unit time. p is proportional to the diffusion coefficient for diffusion into the material near the surface and inversely proportional to the thickness, h, of layer in which the concentration gradient is set up. That is,

$$p = \frac{D}{h} = \frac{D_0}{h} \exp(-E/kT), \qquad \ldots\ldots(2)$$

$$\alpha = \sigma_{\mathrm{ion}}/\sigma_{\mathrm{el}}. \qquad \ldots\ldots(3)$$

It is to be noted that the diffusion may not take place entirely through crystals, but to some extent along boundaries between crystals.

---

* The electron-microscope studies described by Ahearn and Becker refer to thoriated tungsten. They show that thorium comes to the surface in "eruptions" at a relatively small number of randomly located points.

The thermionic emission is related to the number of adatoms by Langmuir's equation

$$j = j_0 \,.\, e^{-a(N_s - N)}. \qquad \ldots\ldots(4)$$

Eliminating N from (1) and (4) gives the equation of current,

$$\frac{1}{j}\frac{dj}{dt} = p \log\left(\frac{j_0}{j}\right) - \frac{(a\alpha)}{e}\, j. \qquad \ldots\ldots(5)$$

Its approximate solution is found by Sproull (*loc. cit.*) to be

$$\frac{j}{j_1} = 1 + \frac{2}{Ae^{qt} - 1}, \qquad \ldots\ldots(6)$$

where j_1 is the asymptotic value of the current for large values of t. It is found from equation (5) on putting $dj/dt = 0$. Thus

$$p \log\left(\frac{j_0}{j_1}\right) = \left(\frac{a\alpha}{e}\right) j_1. \qquad \ldots\ldots(7)$$

The decay constant q is found to be

$$q = \frac{\dfrac{2a\alpha}{e} j_1}{1 - \exp\left(-\dfrac{2a\alpha}{pe} j_1\right)}. \qquad \ldots\ldots(8)$$

This is slightly different from the expression found by Sproull and is more accurate when $\dfrac{a\alpha}{pe} j_1$ is small.

A is found from the boundary condition, that at time zero the number of adatoms in excess of N_1 (the number corresponding to j_1) is n. This gives

$$A = \frac{e^{an} + 1}{e^{an} - 1}, \qquad \ldots\ldots(9)$$

since

$$j(0) = j_1 e^{an}. \qquad \ldots\ldots(10)$$

We shall now use these results to derive an expression for the spectrum of flicker noise on the assumption that variations in emission arise from eruptions of barium atoms on to the surface at scattered points and that decay of emission from an eruption is controlled by the diffusion-conduction process described above.

§ 3. FLICKER EFFECT

Our problem essentially is to calculate the correlation function $f(\tau)$ of the noise current from an emitting patch. For the spectral power density $S(\omega)$ is the Fourier cosine transform of $f(\tau)$. Thus (Wang and Uhlenbuk, 1945)

$$S(\omega) = 4\int_0^\infty f(\tau)\cos(\omega\tau)\,d\tau. \qquad \ldots\ldots(11)$$

Now j_1 is the mean current and N_1 is the mean surface density of polarized adatoms. The increase in current due to $n + N_1$ adatoms is

$$\Delta j = j_1(e^{an} - 1). \qquad \ldots\ldots(12)$$

The correlation function of this noise current is, then,

$$\begin{aligned} f(\tau) &= \langle \Delta j(t) \,.\, \Delta j(t+\tau) \rangle_{\text{Av}}, \\ &= j_1^2 \langle (e^{an(t)} - 1)(e^{an(t+\tau)} - 1) \rangle_{\text{Av}}, \end{aligned} \qquad \ldots\ldots(13)$$

where the mean is to be taken over all possible values of n and t. To evaluate this mean we need to know the distribution law of lifetime of adatoms on the surface. That is to say, if we have n excess adatoms at time zero, how many of these would one expect to find after time τ? But this we have already calculated in § 3. Thus from equations (6) and (10)

$$\frac{<n(\tau)>_{\text{AV}}}{n(0)}=\frac{\log\left(1+\dfrac{2}{Ae^{q\tau}-1}\right)}{\log\left(1+\dfrac{2}{A-1}\right)}. \qquad \ldots\ldots(14)$$

Therefore in equation (12) the mean value for a given n is

$$j_1^2(e^{an}-1)^2\,\frac{<e^{an(\tau)}-1>_{\text{AV}}}{e^{an(0)}-1}=j_1^2(e^{an}-1)^2\,.\,\frac{A-1}{Ae^{q\tau}-1}. \qquad \ldots\ldots(15)$$

Finally, to find $f(\tau)$ we have to average over all possible values of n. This is where a knowledge of the distribution law $W(n)$ of excess adatoms enters. We have already pointed out that the standard deviation for n is not the square root of N_1 but may be many times greater. Let us therefore leave it to be found from comparision of our final expression for $S(\omega)$ with experiment and denote it by N. In the absence of any exact knowledge about $W(n)$ we shall take it to be Gaussian. Thus

$$W(n)=\frac{1}{N\sqrt{(2\pi)}}\exp\left(\frac{-n^2}{2N^2}\right). \qquad \ldots\ldots(16)$$

Then

$$f(\tau)=j_1^2\int_{-\infty}^{+\infty}\frac{A-1}{Ae^{q\tau}-1}(e^{an}-1)^2\,.\,W(n)dn, \qquad \ldots\ldots(17)$$

where A is the function of (an) given by equation (9).

Write

$$g(\tau,\,an)=\frac{A-1}{Ae^{q\tau}-1}, \qquad \ldots\ldots(18)$$

then equation (17) becomes

$$f(\tau)=\frac{j_1^2}{a}\int_{-\infty}^{+\infty}g(\tau,\,n)(e^n-1)^2\,.\,W\left(\frac{n}{a}\right)dn. \qquad \ldots\ldots(19)$$

This integral can be readily evaluated approximately since $g(\tau,\,n)$ is a slowly, and almost linearly, varying function of n. Thus

$$\begin{aligned}f(\tau)&=\frac{j_1^2}{a}\frac{1}{N\sqrt{(2\pi)}}\int_{-\infty}^{+\infty}g(\tau,\,n)(e^{2n}-2e^n+1)\exp\left(\frac{-n^2}{2a^2N^2}\right)dn,\\&\simeq j_1^2[e^{2v}\,.\,g(\tau,\,2v)-2e^{\frac{1}{2}v}\,.\,g(\tau,\,v)+g(\tau,\,0)],\end{aligned} \qquad \ldots\ldots(20)$$

where

$$v=(aN)^2. \qquad \ldots\ldots(21)$$

The spectral power density is now obtained from equation (11). Although the Fourier transform of $g(\tau,\,v)$ could be evaluated, probably in series form, the following procedure leads to a more simple and illuminating formula for $S(\omega)$. It depends on the observation that the function $f(\tau)$, given by equation (20), is of the same form as

$$f_m(\tau)=c\,.\,(q\tau)^mK_m(q\tau). \qquad \ldots\ldots(22)$$

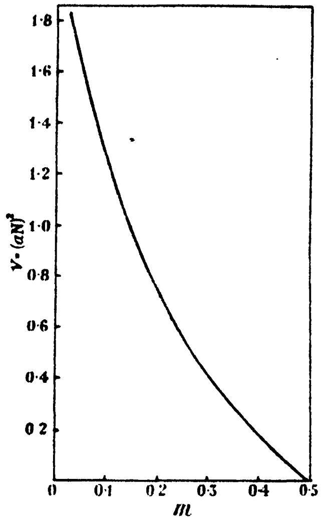
where $K_m(x)$ is the modified Bessel function of order m. Both fall off exponentially for large values of τ and rise more rapidly than $\exp(-q\tau)$ as τ tends to 0. The constant c is found by evaluating $f(0)$ and $f_m(0)$. This gives

$$c = j_1{}^2(e^{2v} - 2e^{\frac{1}{2}v} + 1)\frac{2^{1-m}}{\Gamma(m)}. \qquad \ldots\ldots(23)$$

The order m is found in terms of v by equating the integrals of $f(\tau)$ and $f_m(\tau)$ over the infinite range $0 < \tau < \infty$. This gives

$$\frac{\Gamma(\frac{1}{2})\Gamma(m+\frac{1}{2})}{\Gamma(m)} = \frac{\dfrac{2e^{2v}}{e^{2v}-1}\log\left(\dfrac{e^{2v}+1}{2}\right) - \dfrac{4e^{\frac{1}{2}v}}{e^{v}-1}\log\left(\dfrac{e^{v}+1}{2}\right) + 1}{e^{2v} - 2e^{\frac{1}{2}v} + 1}. \qquad \ldots\ldots(24)$$

m is shown as a function of $v = (aN)^2$ in the graph. It shows that as v decreases from ∞ to 0, m increases from 0 to 0·5.



Dependence of $(aN)^2$ on m.

Returning to the correlation function given by equations (22) and (23), we can now evaluate the spectral power density in closed form. Thus

$$\begin{aligned} S(\omega) &= 4\int_0^\infty f_m(\tau)\cos(\omega\tau)d\tau \\ &= 4j_1{}^2(e^{2v} - 2e^{\frac{1}{2}v} + 1)\frac{2^{1-m}}{\Gamma(m)}\int_0^\infty (q\tau)^m \,.\, K_m(q\tau)\cos(\omega\tau)d\tau \\ &= 4j_1{}^2(e^{2v} - 2e^{\frac{1}{2}v} + 1)\frac{\Gamma(\frac{1}{2})\Gamma(m+\frac{1}{2})}{\Gamma(m)}\cdot\frac{q^{2m}}{(\omega^2+q^2)^{m+\frac{1}{2}}}. \qquad \ldots\ldots(25) \end{aligned}$$

Now $S(\omega)$ given by equation (25) is the spectral power density of emission from one eruption of barium atoms on to the surface. j_1 is the average current-density of emission from patches at eruptions. But over the entire emitting surface the average fraction of area affected by eruptions is small; denote it by γ. Since eruptions are uncorrelated, the spectral power density of the noise current from the whole surface is

$$\overline{\Delta j^2} = \gamma \,.\, S(\omega). \qquad \ldots\ldots(26)$$

§4. DISCUSSION OF RESULTS

(i) The dependence of flicker noise on current and frequency is obtained from equation (25) on substituting for q from equation (8). It is convenient to normalize the current j and denote it by i, thus

$$i=\frac{2a\alpha}{pe}\,j_1. \qquad \ldots\ldots(27)$$

Then

$$\overline{\Delta j^2}=4\gamma\left(\frac{pe}{2a\alpha}\right)^2.\,p^{2m}(e^{2v}-2e^{\frac{1}{2}v}+1)\,\frac{\Gamma(\frac{1}{2})\Gamma(m+\frac{1}{2})}{\Gamma(m)}\,\frac{i^{2m+2}}{(1-e^{-i})^{2m}}\,\frac{1}{(\omega^2+q^2)^{m+\frac{1}{2}}},$$

$$\propto\frac{i^{2m+2}}{(1-e^{-i})^{2m}}\,\frac{1}{\omega^{2m+1}}\quad\text{for }\omega\gg q, \qquad \ldots\ldots(28.1)$$

$$\propto\frac{i^{2m+2}}{\omega^{2m+1}}\quad\text{for }\omega\gg q\text{ and }i\gg 1. \qquad \ldots\ldots(28.2)$$

Therefore for large currents i and for frequencies greater than q the index of $1/\omega$ exceeds the index of i by unity. This simple relationship is therefore to be looked for in experimental results.

We have developed the above theory from the model of a thermionic emitter in which emission occurs in patches over the surface. Each patch is probably about the size of the exposed area of a crystal. This suggests that flicker noise will occur at the surface of other crystals emitting electrons, provided these crystals contain mobile polarizable impurity centres which can diffuse on to the surface and thereby alter the effective work function. It is therefore most likely to be found in multi-crystalline masses of impurity semi-conductors, of which photo-conductive lead sulphide is an example. γ is then the average fractional number of contacts at which eruptions of foreign atoms on to the contact surface occur.

In table 2 of the paper by Harris, Abson and Roberts, measured values of the current index, denoted by x, and the frequency index, y, are tabulated for lead-sulphide cells and carbon resistors. When account is taken of the scatter in points from which these index values were obtained, the agreement with the law $y=x+1$ predicted on the above theory is satisfactory. In the case of the carbon resistors, the agreement is especially close.

(ii) The magnitude of the decay constant q and its dependence on temperature are given by equations (2) and (18). Thus

$$q=p\frac{i}{1-e^{-i}},$$

$$p=\frac{D_0}{h}\exp(-E/kT),$$

and

$$i=\frac{2a\alpha}{pe}j_1. \qquad \ldots\ldots(28.3)$$

The normalized current i can also be expressed in terms of the current ratio j_0/j_1 of Sproull's experiment (1945). Thus

$$i=2\log(j_0/j_1).$$

Therefore

$$\frac{q}{p}=\frac{2\log(j_0/j_1)}{1-(j_1/j_0)^2}.$$

Thus as j_0/j_1 increases from 1 to 100, q/p only increases from 1 to 10. Moreover, the experiments of Sproull have shown that for an oxide-coated filament (50% BaO and 50% SrO) in the temperature range 950° K. to 1100° K., j_0/j_1 is around 10. q and p are therefore of the same order of magnitude,

$$q \sim p = \frac{D_0}{h} \exp(-E/kT).$$

From values of q at three temperatures, deducible from Sproull's results, one gets

T(°K.)	950	1020	1100
q(sec^{-1})	2500	6100	20000

The value of E is therefore about 1·23 ev. and

$$\frac{D_0}{h} \sim 8 \cdot 10^9.$$

The spectral density of the noise current should therefore flatten out to a constant value as the frequency is decreased below about 400 cycles per second at 950° K. The time constant $1/q$ is therefore of the order of hours at normal air temperatures.

In the case of other materials such as lead sulphide, the value of q depends most markedly on the value of the activation energy. This may lie between 0·5 and 3 ev. If we take $D_0/h = 10^{10}$ we find that $1/q > 20$ sec. at normal air temperature. This estimate is probably close enough to show that at normal air temperture the time constant of the diffusion-conduction process is of the order of minutes at the least. This explains why Harris, Abson and Roberts were unable to detect any flattening of the noise spectrum at frequencies as low as 1 c.p.s.

(iii) Schottky's formula for flicker noise in valves is a limiting case of our theory, viz. $N = \surd N_1$, $\gamma = 1$, and $\alpha = 0$. Since aN_1 is less than 100 and $N_1 \sim 10^{14}$, aN is a very small quantity. Then $m = 0{\cdot}5$, and one gets

$$\left.\begin{aligned} \overline{\Delta j^2} &= 4 j_1{}^2 a^2 N_1 \frac{q}{\omega^2 + q^2}, \\ &\propto j_1{}^2/\omega^2 \text{ for } \omega \gg q. \end{aligned}\right\} \qquad \ldots\ldots (28)$$

In addition, $\alpha = 0$ means there is no ionic current and that $q = p$.

On our theory, therefore, departure of the spectral density law from $1/\omega^2$ for $\omega \gg q$ is an indication of the width of the distribution of n, whereas the departure of the current law from j^2 indicates the presence of ionic current. Schottky attempted to fit equation (28) to the spectral distribution measured by Johnson for an oxide-coated filament. The fit was not too good. The law observed was more nearly $S(\omega) \propto 1/\omega^{1{\cdot}25}$. This would correspond to $m = 0{\cdot}125$ and $(aN)^2 = 1{\cdot}11$. From equation (28.2) this flicker noise should depend on current as $j^{2{\cdot}25}$, assuming ionic current. This is in good agreement with Johnson's findings for large currents.

(iv) In order to account for the deviation of the frequency dependence of the noise power density from $1/\omega^2$ it seems necessary to postulate that impurity centres diffuse on to the emitting surface of crystals in clusters. This leads one to speculate on mechanisms that could account for the continual production and migration of clusters of atoms through a semi-conductor. A clue to a possible mechanism is to be found in the theory of photolysis of AgBr under the action of light, as proposed by Mott and Gurney (1940). To be specific let us consider the case of the BaO emitter.

The process of activation is designed to dissociate some of the BaO into free barium and to drive off the liberated oxygen. Thus in an active emitter one finds a considerable excess of barium. This free barium will most likely occur in the form of colloidal specks of barium atoms, since the process of dissociation will not be equally likely to occur at any point in the original mass of BaO, but will be more likely to commence and grow from points where the crystal lattice is distorted as at interstitial Ba ions. Moreover, specks will only be found in cracks or on the surface of crystals, for there would not be enough room for them in the body of a crystal. Now these specks of Ba will be scattered at random throughout a mass of activated BaO. However, specks which are at the exposed surface of a crystal, particularly the emitting surface of the mass of oxide, will in general be in a strong electric field, Some of the atoms in such specks will lose valency electrons, and those that remain as atoms will be strongly polarized. The loosely bound positive ions will be repelled. In fact, the speck will tend to diffuse over the entire surface of the crystal and on to adjacent crystals, forming a coating partly of Ba atoms and partly Ba ions. A selected Ba atom will alternate from the atomic to the ionic conditions. Now when the oxide is in an electric field and electrons are being drawn away from the emitting surface, ions of Ba will be conducted away from the surface. These ions will conduct through the body of crystals as well as along cracks. Eventually they will be attracted to specks of Ba. We imagine this attraction to be due to the presence of a negative charge on a speck occasioned by the capture of conduction electrons by the metallic speck, just as photo-electrons are captured by specks of Ag in the photolysis of AgBr (*vide* Mott and Gurney, p. 230). Specks of Ba will therefore grow by capture of electrons and Ba ions. Now the disintegration of a speck of Ba on the emitting surface of a crystal gives rise to a deficit of Ba on the surface. New specks of barium will therefore migrate on to the surface. This migration will be in part due to the concentration gradient near the surface and in part to conduction of specks in the electric field, since they are on the average negatively charged. Thus we have a complete cycle that accounts for the growth of specks of barium atoms, their migration to the emitting surface of the oxide, their disintegration into Ba ions at the surface, conduction of these ions away from the surface and their recombination with negatively charged specks of barium atoms. The only question outstanding is how new specks of barium are formed. Mott and Gurney encountered the same problem for specks of Ag in AgBr and explained it in terms of surface irregularities or "reifkeime". Perhaps a similar explanation is possible in this case. On the other hand we are probably safe in assuming that some of the Ba ions will conduct right through the oxide and deposit barium at the core of the emitter. Thus a layer of barium will be formed at the core and this will diffuse into the oxide along cracks and provide the nuclei of atomic barium we need.

ACKNOWLEDGMENTS

In conclusion, the author desires to express his thanks to Dr. R. A. Smith of the Telecommunications Research Establishment for his helpful criticism and encouragement, and to the Ministry of Supply for permission to publish this paper.

REFERENCES

Ahearn, A. J. and Becker, J. A., 1938. *Phys. Rev.*, **54**, 448.
Benjamin, M., Huck, R. J. and Jenkins, R. O., 1938. *Proc. Phys. Soc.*, **50**, 345.
Blewitt, J. P., 1939. *Phys. Rev.*, **55**, 713.
Christensen and Pearson, 1936. *B.S.T.J.*, **15**, 197.
Harris, E. J., Abson, W. and Roberts, W. L., 1947. (In preparation.)
Johnson, J. R., 1925. *Phys. Rev.*, **26**, 71.
Miller, Lewis, Schiff and Stephens, 1946. *Phys. Rev.*, **69**, 682.
Mott and Gurney, 1940. *Electronic Processes in Ionic Crystals* (Cambridge: The University Press).
Schottky, W., 1926. *Phys. Rev.*, **28**, 74.
Schottky and Rothe, 1928. *Handb. Experimentalphys.*, **13**, 276.
Sproull, R. L., 1945. *Phys. Rev.*, **67**, 166.
Wang, M. C. and Uhlenbuk, G. E., 1945. *Rev. Mod. Phys.*, **17**, 326.

SPONTANEOUS FLUCTUATIONS OF ELECTRICITY IN THERMIONIC VALVES UNDER RETARDING FIELD CONDITIONS

By D. K. C. MacDONALD* and R. FÜRTH†

* Military College of Science; now at Clarendon Laboratory, Oxford
† Department of Mathematical Physics, University of Edinburgh; now at Birkbeck College, University of London

MS. received 1 *October* 1946 ; *read* 21 *February* 1947

ABSTRACT. It follows from theoretical considerations that the shot effect in diodes should satisfy the classical formula of Schottky under extreme retarding-field conditions when the product IR of current and differential resistance of the valve reaches the constant value kT/e. It appears from the results of the experiments described in the present paper, that both relations are satisfied for currents not exceeding a certain critical value I_c. A theory is presented which permits calculation of I_c for a given valve structure. Finally, it is shown how measurements of this type can be used satisfactorily for determining the cathode temperatures in diodes.

§ 1. INTRODUCTION

It was first pointed out by W. Schottky (1918) in a classical paper that, owing to the electronic structure of electricity, the electric current through a thermionic valve should exhibit irregular fluctuations. This phenomenon is usually called the *Shot Effect*. If it is assumed that the emission of the individual electrons from the cathode is a sequence of random events and that the transits of these electrons through the valve are independent of each other, one can derive a formula for the mean square fluctuation

$$\overline{(\delta i)^2} = \overline{(i-\bar{i})^2}$$

of the current i. Considering these fluctuations as a superposition of harmonic oscillations with frequencies f and random phases, one obtains for the part $\overline{(\delta i)_f^2}\,df$ of $\overline{(\delta i)^2}$, which is due to oscillations in a narrow region between f and $f+df$ of the spectrum

$$\overline{(\delta i)_f^2}\,df = 2eI\,df, \qquad \ldots\ldots(1)$$

where e is the magnitude of the electronic charge and $I=\bar{i}$ the average current.

If the fluctuations are observed by means of an instrument (plus attached network) having a response $\phi(f)$ for an alternating current of frequency f, then the total effective fluctuation of current will be given by

$$\overline{(\delta i)^2}=2eI\int_0^\infty \phi(f)\,df=2eI\Delta f, \qquad \ldots\ldots(2)$$

where Δf is the integrated band-width of the instrument with its associated circuit.

The formulae (1) and (2) apply in the case of a valve in a state of saturation where the above-mentioned conditions are satisfied, and experiments carried out under these conditions have completely verified the theory (Williams *et al.*, 1926 and 1929). Under the ordinary working conditions of a valve, however, a negative space charge is established in the inter-electrode space which produces a variable barrier of a certain mean height for the transit of the electrons. As the height of the barrier increases with increasing current, one can easily see that the magnitude of the fluctuations will be less than the value given by formula (2), or

$$\overline{(\delta i)^2}=2eI\Gamma^2\Delta f \quad 0<\Gamma^2\leqslant 1, \qquad \ldots\ldots(3)$$

where Γ^2, the so-called *space-charge reduction factor*, depends on the operating conditions of the valve. An extensive literature exists on the theoretical evaluation of Γ^2 (e.g. Johnson, 1925; Llewellyn, 1930; Schottky and Spenke, 1937). North and collaborators (1940–1942) in particular have carried out extensive experimental work on the measurement of the shot effect under space-charge conditions, evidencing good agreement with theory.

The above explanation of the space-charge reduction implies that $\Gamma^2=1$ when there is no variable potential barrier within the valve, i.e. when the potential either increases or decreases monotonically from cathode to anode. The first of these conditions is realized when the valve is saturated, the second when the valve is operated under retarding-field conditions, i.e. when the cold electrode has a sufficiently high negative potential with respect to the hot one. The theory therefore suggests that under true retarding-field conditions, Γ^2 should become unity once more and formula (2) should again be applicable.

It is a well-known fact that the phenomenon of emission of electrons from a hot electrode can be explained on the basis of statistical mechanics by assuming that the electrons have to overcome a fixed negative potential barrier v_c at the surface of the metal in the process of emission. Consequently there exists a probability for the escape of an electron through the surface which is proportional to $\exp[ev_0/\boldsymbol{k}T]$ where T is the temperature of the metal, and which is independent of the fate of the other electrons. As this expression is also quite independent of the structure of the barrier, it can equally well be applied in the case of an electronic valve under retarding-field conditions. Thus the probability of the transit of an electron from the interior of the hot electrode to the interior of the cold electrode is proportional to

$$\exp[e/\boldsymbol{k}T.(v_c+v_a+V)],$$

where V is the (negative) anode potential and v_a a possible potential barrier at the surface of the cold electrode. Again, the probabilities for the transit of different electrons will be independent of each other.

It appears that the conditions for the application of the formulae (1) and (2) are satisfied in both above-mentioned cases. Furthermore, the current I under true retarding-field conditions will be connected with the saturation current J by

$$I = J \,.\, \exp\left[\frac{e(V+v_a)}{kT}\right] = C \,.\, \exp\left[eV/kT\right]. \qquad \ldots\ldots(4)$$

It follows immediately from (4) that the differential resistance R of the valve under these conditions satisfies the relation

$$R \equiv \frac{1}{\partial I/\partial V} = \frac{kT}{eI}, \qquad \ldots\ldots(5)$$

showing that the product IR ought to depend on the temperature T. The relation (5) can thus be regarded as the experimental criterion for the establishment of the true retarding-field condition, and the shot effect reduction factor Γ^2 should become unity when this relation is satisfied.

An attempt to measure the shot effect in diodes under retarding-field conditions has been made before (Williams, 1936) but detailed evidence will be given below to show that the retarding region was not in fact entered on that occasion; in particular, equation (5) was not satisfied. Further, these experiments were marred by low-frequency flicker effect (Johnson, 1925; Schottky, 1926), and thus no conclusion could be drawn as to whether the shot-effect fluctuations were properly described by theory. It was therefore decided to conduct a new and more thorough investigation in order to clarify the situation. The main results have been published in a short note (Fürth and MacDonald, 1946) and a detailed account of the work is presented in what follows.

§ 2. THE FLUCTUATION MEASUREMENTS

The standard technique of fluctuation measurement that has been developed in recent years by North was employed again in this investigation. It consists in comparing the fluctuations to be measured with those generated by a diode operating under saturation condition, where equation (2) is known to be precisely satisfied (Williams, 1926 and 1929). If the same amplifier and detector are used for both measurements, the amplification factor and band width of the measuring device drop out of the final formulae and therefore need not be determined.

The circuit diagram is shown in figure 1. It consisted of the fluctuation measurement unit A which was specially designed for this investigation in the laboratory, the pre-amplifier B, the main amplifier C, a final amplifier D, the diode detector E and the measuring instrument F.

The unit A contained the test valve and the comparison valve. The test valves used were all close-spaced diodes of simple cylindrical structure with indirectly heated cathodes; the comparison valve was a directly heated tungsten-filament diode working under temperature-limited conditions. The main amplifier C was a Standard Marconi Receiver type CR.100/2; pre-amplifier and main amplifier were tuned to a frequency range around 1·5 to 2 Mc./s, to minimize interference and to avoid the "flicker effect" (Johnson, 1925; Schottky, 1926). The band width was approximately 6 kc./s. The task of the power amplifier was to provide an average input r.m.s. voltage of the fluctuations to the detector valve of the order of 50 volts; a careful examination proved that under these conditions the detector was perfectly linear within the experimental error. The measuring instrument F was a standard AVO milliammeter, the readings of which were directly proportional to the r.m.s. value of the fluctuations to be measured.

Precautions were of course taken to ensure that the effect of extraneous disturbances was reduced to a minimum throughout. Also, in addition to the meter indication, the fluctuation was displayed visually on a cathode-ray oscillograph and made audible by a loud-speaker; any occasional disturbance was then immediately evident and measurements could be suspended until the disturbance was removed.

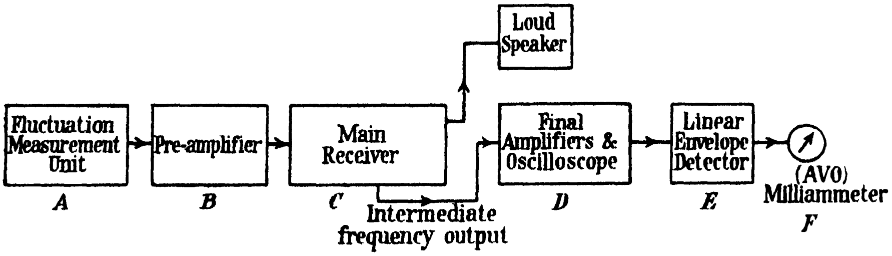

Loud Speaker
Fluctuation Measurement Unit
A
Pre-amplifier
B
Main Receiver
C
Intermediate frequency output
Final Amplifiers & Oscilloscope
D
Linear Envelope Detector
E
(AVO) Milliammeter
F

Figure 1(*a*). Schematic of fluctuation measurement layout.

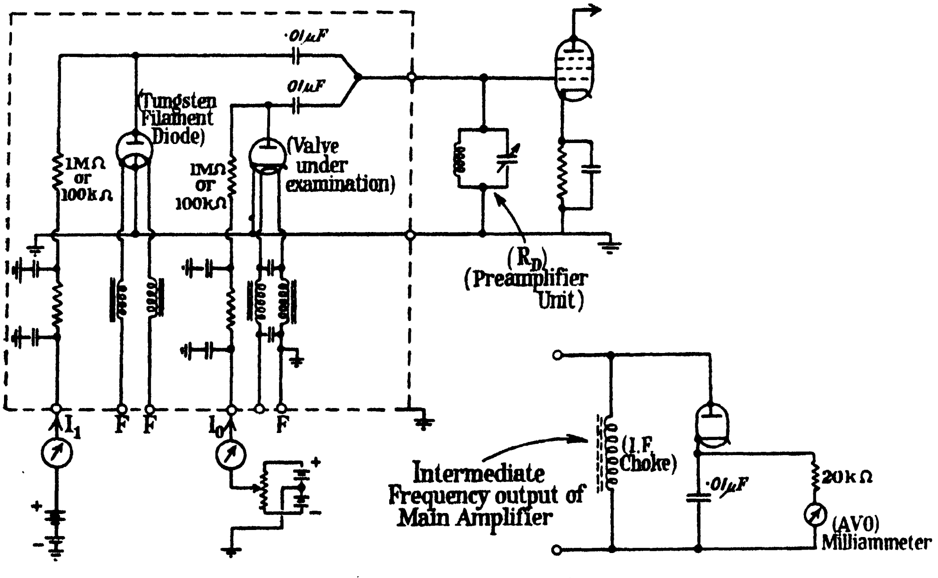


Figure 1(*b*). Fluctuation measurement unit.

Figure 1(*c*). Envelope detector unit.

The following experimental procedure was adopted for the measurements:—First, the input to the amplifiers was short circuited and V_1, the output voltage due to inherent fluctuations introduced by the measuring device, was read. Then the short circuit was removed and the test valve alone operated at the required current I; the output voltage V_2 is then, according to (3), given by

$$V_2^2 = V_1^2 + \rho^2(2eI\Gamma^2 + 4kT_r/R_D), \qquad ......(6)$$

where the third term represents the thermal fluctuations in the first circuit and ρ represents the impedance of the tuned circuit and the test valve in parallel; R_D is the dynamic resistance of that circuit and T_r its temperature. Finally the saturated tungsten diode was introduced and the current J_λ through this valve adjusted (by varying its heater current) until the output voltage V_3, which now satisfies

$$V_3^2 = V_2^2 + \rho^2 . 2eJ, \qquad \ldots\ldots(7)$$

becomes equal to

$$V_3^2 = 2V_2^2 - V_1^2. \qquad \ldots\ldots(8)$$

From (6), (7) and (8) we have for the required reduction factor:

$$\Gamma^2 = \frac{1}{I}\left(J - \frac{2kT_r}{eR_D}\right). \qquad \ldots\ldots(9)$$

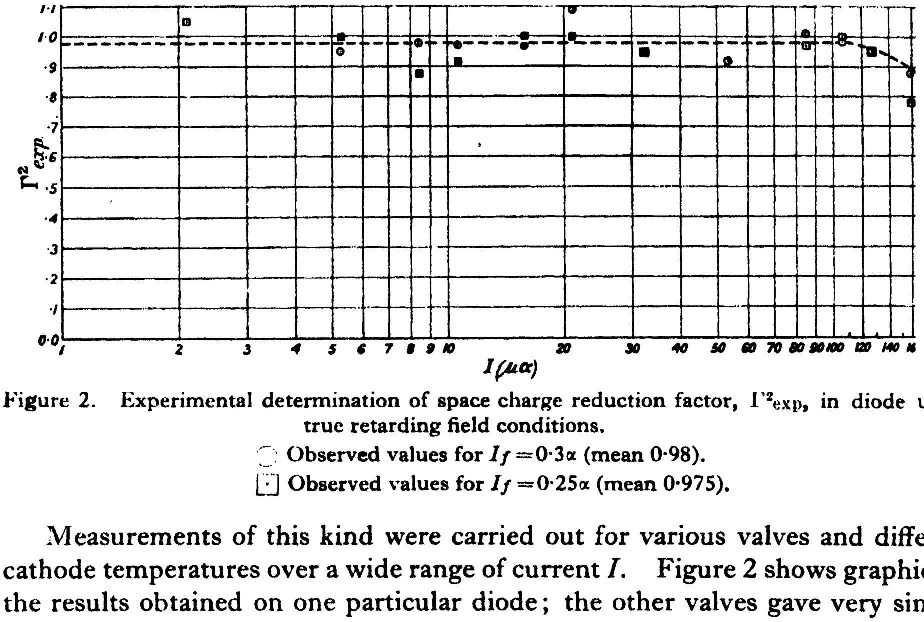


Figure 2. Experimental determination of space charge reduction factor, Γ^2_{exp}, in diode under true retarding field conditions.

○ Observed values for $I_f = 0{\cdot}3\alpha$ (mean 0·98).

⊡ Observed values for $I_f = 0{\cdot}25\alpha$ (mean 0·975).

Measurements of this kind were carried out for various valves and different cathode temperatures over a wide range of current I. Figure 2 shows graphically the results obtained on one particular diode; the other valves gave very similar results. It appears that the space-charge reduction factor is, indeed, unity up to a limiting current I_c, which in the particular case shown in figure 2 is approximately 100 μA.

§ 3. THE MEASUREMENT OF DIFFERENTIAL VALVE RESISTANCES UNDER EXTREME RETARDING-FIELD CONDITIONS

In order to check the validity of formula (5) in the true retarding field region, a method for measuring with precision the differential resistances of the valves used for the fluctuation measurements in relation to the current I had to be developed. A bridge method was employed, the circuit diagram of which is shown in figure 3. The bridge was fed from a beat frequency oscillator and frequencies between 1 kc./s. and 10 kc./s. were used. Balance was observed on a cathode-ray oscillograph after suitable amplification. A large resistance R_3 was used in series with the test valve, bridged by a large condenser C_1, to prevent slow drift of the measured current I over periods of the order of several seconds resulting from small battery variations and slow variations of emission, which latter may be expected to arise in oxide-coated cathodes when operated at low filament-temperatures.

Precautions had to be taken in order to avoid systematic errors arising from the non-linearity of the valve characteristic. As this point is of some importance for all investigations on valves in the extreme retarding-field region, the relevant mathematical treatment is given in some detail.

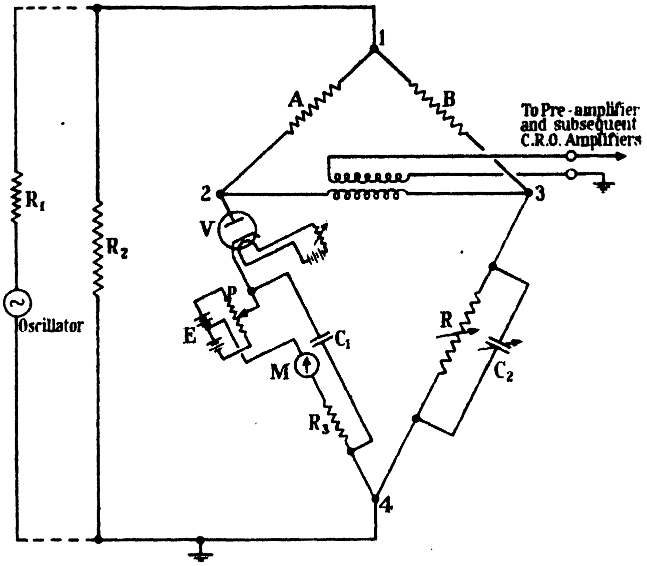


Figure 3. Detail of bridge circuit for differential valve resistance measurement.

V : Valve under examination.
R_1, R_2 : Potentiometer to provide low input voltage to bridge.
A, B : Fixed (resistive) bridge arms.
E, P : Battery and potentiometer to adjust valve current.
R : Variable (resistive) bridge arm.
C_2 : Variable condenser to balance reactive component.

According to (4) the mean current through the valve will, by the application of a small alternating voltage, $v \sin \omega t$, be changed from its original value

$$I_0 = C \, . \exp\,[eV_0/kT] \qquad \ldots\ldots(10)$$

to

$$I = C \, . \exp\,[e/kT(V + v \sin \omega t)], \qquad \ldots\ldots(11)$$

where V, the average anode potential, is different from V_0 because of the voltage drop across the external resistance R' due to the average increase in current. We have clearly

$$\delta \equiv V - V_0 = R'(I_0 - \bar{I}), \qquad \ldots\ldots(12)$$

and thus by means of (10) and (11), up to the second order in v,

$$\delta = -\frac{R'I_0}{4} v^2 \left(\frac{e}{kT}\right)^2 \frac{1}{1 + R'I_0 e/kT}. \qquad \ldots\ldots(13)$$

Now R' was chosen so as to make $R'I_0/kT \gg 1$; accordingly (13) simplifies to

$$\delta = -\frac{v^2 e}{4kT}, \qquad \ldots\ldots(14)$$

and the current I now becomes, to the same accuracy,

$$I=C.\exp\left[e/\boldsymbol{k}T\left(V_0-\frac{v^2e}{4\boldsymbol{k}T}\right)+e/\boldsymbol{k}T.v\sin\omega t\right]$$

$$\approx I_0+\frac{I_0e}{\boldsymbol{k}T}\left[1-\frac{v^2}{4}\left(\frac{e}{\boldsymbol{k}T}\right)^2\right].v\sin\omega t-\frac{I_0}{4}\left(\frac{e}{\boldsymbol{k}T}\right)^2 v^2\cos 2\omega t. \qquad \ldots\ldots(15)$$

Hence it appears that first the average current $\bar{I}$ is equal to I_0, which demonstrates the stabilizing effect of the external resistance R' mentioned above. Further, if the disappearance of the fundamental frequency ω on the cathode-ray oscillograph trace be taken as the criterion for the balance of the bridge, the indicated differential resistance R_i of the valve is equal to the reciprocal of the coefficient of the second term in (15), i.e.

$$R_i=R\left[1-\frac{v^2}{4}\left(\frac{e}{\boldsymbol{k}T}\right)^2\right]^{-1}\sim R\left[1+\frac{v^2}{4}\left(\frac{e}{\boldsymbol{k}T}\right)^2\right], \qquad \ldots\ldots(16)$$

instead of the true value given by (5). Finally, it appears that higher harmonics of the applied oscillation are generated by the valve, the first of which is represented by the last term in (15); these harmonics are, of course, not balanced out and will appear on the oscillograph screen, thus reducing the accuracy of the measurement.

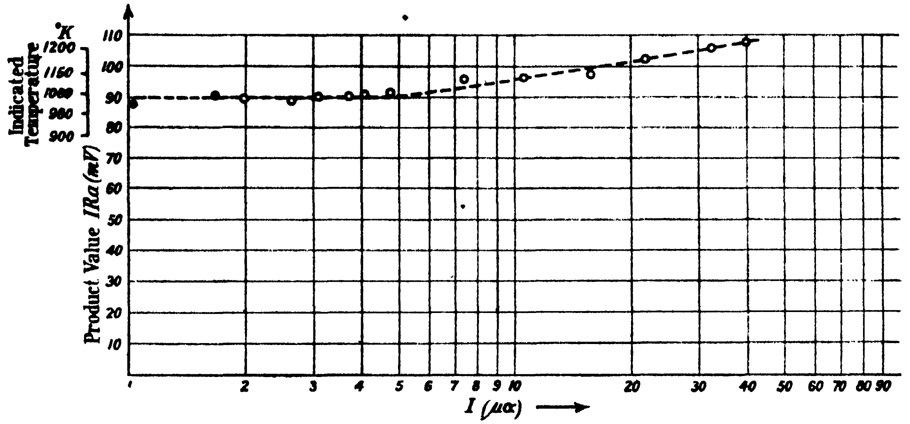


Figure 4. Experimental investigation of valve characteristic in retarding field region—Type 6H6 diode. ($I_f = 0{\cdot}3$a.)

To obtain accurate results one has therefore to restrict the amplitude of the applied voltage v to sufficiently small values for $ve/\boldsymbol{k}T$ to remain below a certain small fraction. Under the present conditions, an accuracy of 1% in the measurement of R required v to be kept in the order of magnitude of 10 mv. To confirm this result, observations were made with inputs ranging from about 2 mv. to 17 mv. and found to be entirely consistent.

In view of the small magnitude of the applied voltage, care had to be taken to ensure that stray voltage pick-up was avoided, which would vitiate the balance; and of course sufficient amplification had to be provided for the cathode-ray oscillograph.

The results obtained on all the valves examined showed clearly that below a certain value I_c of current the product IR became very accurately constant.

Examples of results on two different types of valves are shown in figures 4, 5, and 6. The results represented in figure 4 refer to a diode with a greater electrode spacing than the ones used for the experiments represented by figures 5 and 6, which were also used for the fluctuation experiments described in § 2. The measurements illustrated in figure 5 in particular were carried out on the same valve (under the same operating conditions) as that used for the measurements relating to figure 2. Figure 6 illustrates the effect of the variation of cathode temperature.

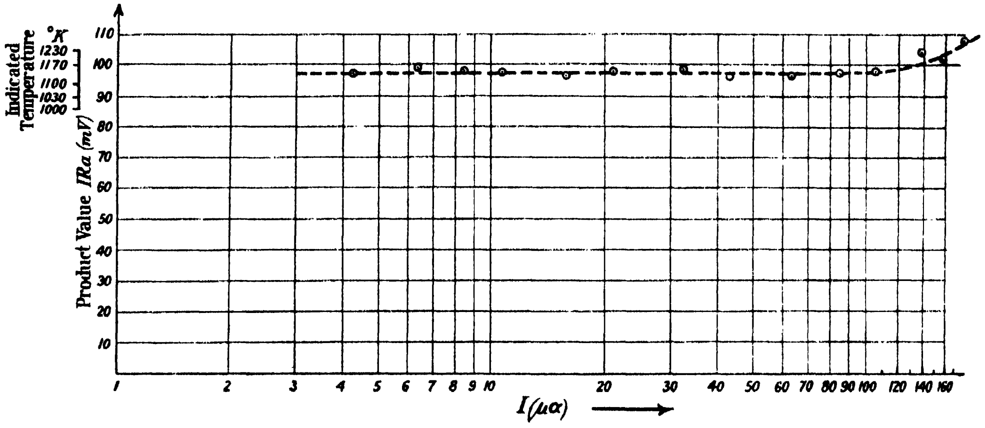


Figure 5. Examination of valve characteristic under retarding field conditions. Close-spaced diode (second model). $(I_f = 0{\cdot}3\alpha\ ;\ V_f = 6{\cdot}9V.)$

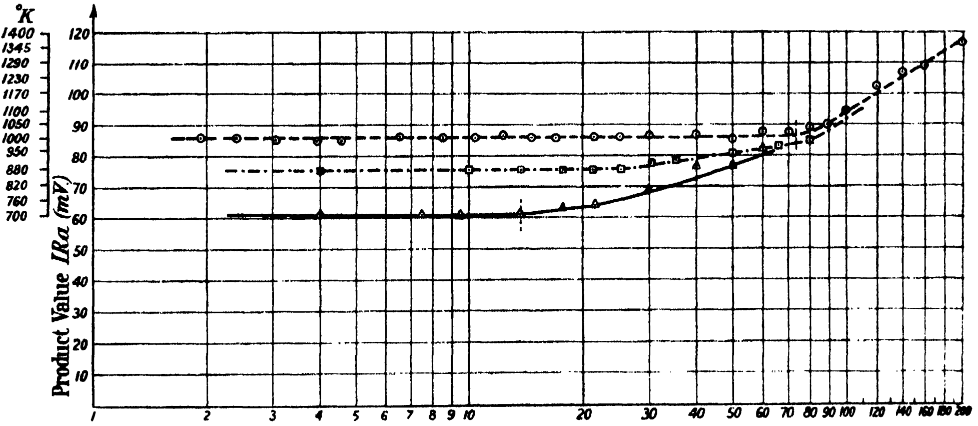


Figure 6. Examination of valve characteristic under retarding field conditions: close-spaced diode (first model).

○ Experimental points 1st series $(I_f = 0{\cdot}3\alpha\ ;\ V_f = 6{\cdot}1V.)$

□ „ „ 2nd series $(I_f = 0{\cdot}25\alpha\ ;\ V_f = 4{\cdot}3V.)$

△ „ „ 3rd series $(I_f = 0{\cdot}2\alpha\ ;\ V_f = 2{\cdot}6_8\ V.)$

The limiting current, as indicated by these graphs, is seen to be much smaller for the widely-spaced diode than for the closely-spaced one, and (for one and the same diode) to increase with increasing cathode temperature. This is in agreement with the theoretical expectations, as will be shown in § 4. One observes also that the limiting current for the application of the fluctuation formula (2) is about the same as that for the application of formula (5) for one and the same valve, in agreement with the theoretical considerations of § 1.

§ 4. THE POTENTIAL DISTRIBUTION IN A DIODE UNDER TRUE RETARDING-FIELD CONDITIONS AND THE CRITICAL CURRENT

The results of the investigations described in §§ 2 and 3 confirm that true retarding-field conditions in a diode will prevail when the current is below a certain critical value. In the following, we attempt to derive a theoretical formula for this critical current which can be directly compared with the experimental results. For this purpose the potential distribution in a diode under retarding field conditions has first to be calculated.

Let us first consider the simplest case of a plane-parallel diode which has been treated by Langmuir (1923) and Fry (1921).

If the cathode with potential $v=0$ is at $x=0$, and the potential minimum of the space-charge barrier with $v=v_m$ is at $x=x_m$, one can, following Langmuir, introduce the dimensionless variables

$$\left.\begin{aligned} \xi &= AT^{-\frac{3}{4}}a^{-\frac{1}{2}}I^{\frac{1}{2}}(x-x_m), \\ \eta &= \frac{e}{kT}(v-v_m), \end{aligned}\right\} \qquad \ldots\ldots(17)$$

where $A=4\left(\frac{\pi}{2k}\right)^{\frac{3}{4}}m^{\frac{1}{4}}e^{-\frac{1}{2}}$, $m=$ electronic mass, and a the electrode surface area. The function $\eta(\xi)$ has been numerically computed by Langmuir, and it appears that $\xi \approx -2{\cdot}5$ for $\eta \geqslant 6$.

As there should be no potential minimum between the electrodes under the true retarding-field regime, the limiting current for this regime is given by $v_m = V_1$, $x_m = d$, where V is the anode potential and d the distance between the electrodes. Hence for $x=0$

$$\xi_0 = -AT^{-\frac{3}{4}}a^{-\frac{1}{2}}I^{\frac{1}{2}}d, \quad \eta_0 = -\frac{e}{kT}V. \qquad \ldots\ldots(18)$$

For an anode potential of $-1/2$ volt and a cathode temperature of 1000° K. one has $\eta_0 \sim 6$, and thus $\xi_0 \sim -2{\cdot}5$; hence for anode potentials larger in absolute value, the critical current I_c becomes

$$I_c \approx \left(\frac{2{\cdot}5}{A}\right)^2 \frac{T^{\frac{3}{2}}a}{d^2} = \frac{(2{\cdot}5)^2\sqrt{2}}{(4\pi)^{\frac{3}{2}}}\,\frac{k^{\frac{3}{2}}}{e\sqrt{m}}\,\frac{T^{\frac{3}{2}}a}{d^2} = 7{\cdot}5\times10^{-6}\frac{T^{\frac{3}{2}}a}{d^2}\ \mu\text{a}. \qquad \ldots\ldots(19)$$

The same problem can be treated in a much simpler way under the condition that the current I is so small that the electrons in the space-charge cloud are practically in thermodynamic equilibrium with the hot cathode of temperature T, or in other words, that the number of electrons constituting the current is small compared with the total number of electrons in the space-charge cloud.

Under this condition the (absolute) space-charge density satisfies the equation

$$\rho = \rho_0 \exp\left[\frac{ev}{kT}\right], \qquad \ldots\ldots(20)$$

which follows directly from statistical thermodynamics. If this is combined with Poisson's equation,

$$\nabla^2 v = 4\pi\rho, \qquad \ldots\ldots(21)$$

one obtains the following differential equation for $v(x, y, z)$:

$$\nabla^2 v = 4\pi\rho_0 \cdot \exp\left[\frac{ev}{\boldsymbol{k}T}\right], \qquad \ldots\ldots(22)$$

which can be solved for given boundary conditions. The density distribution $\rho(x, y, z)$ can be obtained from (20).

This problem has been extensively dealt with by Laue (1918) in a classical paper. For the linear case in question, the solution can be written in the form

$$\rho = \frac{C^2}{8\pi}\frac{e}{\boldsymbol{k}T}\frac{1}{\sin^2\left[\frac{Ce}{2\boldsymbol{k}T}(x+X)\right]}, \qquad \ldots\ldots(23)$$

where C, X are constants of integration, and C has to be real for a distribution with a maximum ρ_m of ρ at a plane $x = x_m$. For x_m and ρ_m, we have at once from (23)

$$x_m = \frac{\pi\boldsymbol{k}T}{Ce} - X, \qquad \ldots\ldots(24)$$

$$\rho_m = \frac{C^2 e.}{8\pi\boldsymbol{k}T}. \qquad \ldots\ldots(25)$$

Supposing $v = 0$, $\rho = \rho_0$ for $x = 0$ at the cathode, we have further,

$$\rho_m = \rho_0 \sin^2\left(\frac{CeX}{2\boldsymbol{k}T}\right). \qquad \ldots\ldots(26)$$

The limit of the true retarding-field regime is given by $x_m = d$ when certainly $\rho_m \ll \rho_0$. Thus from (26),

$$CeX/2\boldsymbol{k}T \ll 1,$$

and consequently, according to (24),

$$C \approx \frac{\pi\boldsymbol{k}T}{de}; \qquad \ldots\ldots(27)$$

and hence from (25),

$$\rho_m = \frac{\pi\boldsymbol{k}T}{8ed^2}. \qquad \ldots\ldots(28)$$

Now from formulae (4) and (20) we obtain for the critical current ,

$$I_c = \frac{J\rho_m}{\rho_0}\exp\left[\frac{ev_a}{\boldsymbol{k}T}\right]; \qquad \ldots\ldots(29)$$

combining this with the classical Richardson expression for the saturation current J,

$$J = \rho_0 a\sqrt{\frac{\boldsymbol{k}T}{2\pi m}}, \qquad \ldots\ldots(30)$$

one obtains finally, with the help of (28),

$$I_c = \frac{\sqrt{\pi}}{8\sqrt{2}}\frac{\boldsymbol{k}^{\frac{3}{2}}}{e\sqrt{m}}\frac{T^{\frac{3}{2}}a}{d^2}\exp\left[\frac{ev_a}{\boldsymbol{k}T}\right] = 6\times 10^{-6}\,\frac{T^{\frac{3}{2}}a}{d^2}\exp\left[-\frac{11\,400}{T}\,|v_a|\,(\text{volt})\right]\mu\text{a}. \qquad \ldots\ldots(31)$$

Apart from the inclusion of the additional exponential factor to allow for the existence of a possible potential barrier at the anode surface to be overcome by the electrons, we observe the agreement of (31) and (19) with a small difference in the numerical factor.

The analogous problem for the more important case of a cylindrical diode can then be treated by the same method, by solving the differential equation (22) for cylindrical symmetry, this proving extremely intractable by the first method (e.g. Wheatcroft, 1940). According to Laue, the dependence of ρ on the distance r from the centre of the cylinders is now given by

$$\rho=\frac{B^2}{2\pi e/\boldsymbol{k}T.r^2\sin^2[B\log(r/R)]}, \qquad \ldots\ldots(32)$$

where B, R are integration constants. It can be readily shown that this formula reduces to (23) within distances δr which are small compared with r, i.e. the plane solution can be also applied to close-spaced cylindrical diodes.

In general, the condition for a maximum ρ_m of ρ at $r=r_m$ is

$$\tan\left(B\log\frac{r_m}{R}\right)=-B, \qquad \ldots\ldots(33)$$

leading to
$$\rho_m=\frac{1+B^2}{2\pi e/\boldsymbol{k}T\,.\,r_m^2}, \qquad \ldots\ldots(34)$$

In the limiting case we have again the maximum appearing at the anode where $r=r_a$, and it can be shown in a manner exactly analogous to that used in the planar case that R can be put very nearly equal to r_c, the radius of the cathode cylinder. Hence from (33),

$$\frac{\tan[B\log(r_a/r_c)]}{B\log(r_a/r_c)}=-\frac{1}{\log(r_a/r_c)}, \qquad \ldots\ldots(35)$$

From this equation B can be obtained for given values of r_a, r_c by means of a table of the function $\tan x/x$ (see e.g. Jahnke–Emde: Funktiontafeln). From (29), (30) and (34) we finally obtain for the limiting current, remembering that now $a=2\pi lr_c$ (l : lengths of anode and cathode cylinders, $l\gg r_a$, r_c),

$$I_c=\frac{1+B^2}{\sqrt{2\pi}}\frac{\boldsymbol{k}^{\frac{3}{2}}}{e\sqrt{m}}\frac{T^{\frac{3}{2}}lr_c}{r_a{}^2}\exp\left[\frac{ev_a}{\boldsymbol{k}T}\right]=15\times10^{-6}(1+B^2)\frac{T^{\frac{3}{2}}lr_c}{r_a{}^2}$$
$$\exp\left[-\frac{11\,400}{T}\,|v_a|\,(\text{volt})\right]\mu\text{a}. \qquad \ldots\ldots(36)$$

When r_a/r_c is of the order of magnitude unity, the right-hand side of (35) becomes large and thus

$$B\approx\frac{\pi}{2\log(r_a/r_c)}\gg1,$$

so that (36) goes over into

$$I_c=\frac{\pi^{\frac{3}{2}}}{4\sqrt{2}}\frac{\boldsymbol{k}^{\frac{3}{2}}}{e\sqrt{m}}\frac{T^{\frac{3}{2}}lr_c}{r_a{}^2\log^2(r_a/r_c)}\exp\left[\frac{ev_a}{\boldsymbol{k}T}\right]. \qquad \ldots\ldots(37)$$

This expression is, apart from the factor r_c/r_a and the exponential factor involving v_a, identical with a formula obtained by Möller and Detels (1926) who used the Langmuir method under certain simplifying assumptions which essentially restrict its application to cases where the treatment given above is valid.

When r_a/r_c is large compared with unity, the right-hand side of (35) becomes small and

$$B \approx \frac{\pi}{\log(r_a/r_c)} \ll 1;$$

hence the term B^2 in (36) can be neglected.

The preceding theoretical results can now be used for calculating the expected values of I_c for the valves used in the present investigation. We restrict ourselves to those measurements where the cathode temperature was known to be the standard temperature of 1000° K. (see also § 5). The other data are collected in the following table:

Valve type	r_c (appr.) (cm.)	r_a (cm.)	l (cm.)	I_c(theor.) ($v_a = 0$) (μa.)	I_c (obs.) (μa.)	v_a (calc.) (v.)
6H6	0·06	0·1	0·6	24	$\sim$5	0·14
CV140	0·06	0·078	0·6	124	$\sim$100	0·02

The theoretical values of I_c, calculated from (36) for $v_a = 0$ are shown in column 5 of this table, and the observed ones in column 6. One sees immediately that I_c should have a much smaller value for the relatively wide-spaced valve 6H6 than for the very close-spaced valve CV140 (we are indebted to Messrs. Ferranti, Ltd., Hollinwood, Manchester, for supplying specimens of this latter valve for this work) which agrees with experiment. This fact demonstrates clearly the advantage of using valves of the latter type for experiments on electrical fluctuations in the retarding-field region.

It is further seen that the observed values of I_c are always smaller than the theoretical ones. This fact has been observed by previous workers in this field (e.g. Möller and Detels, 1926; North, 1940) and several possible explanations have been considered by the present authors. A careful discussion of these possibilities shows, however, that none of these explanations is in fact acceptable in the present work. On the other hand, the theory presented above yields a completely natural explanation when it is assumed that v_a has a small negative value, i.e. if one assumes that not all electrons impinging upon the anode can penetrate the surface. The values of the potential barrier required for restoring agreement between theory and experiment are shown in the last column of table 1. They are of the order of magnitude of 0·1 volt and may therefore be easily accounted for by contamination of the anode surface. The effect is therefore a spurious one; it was indeed, found, to vary from valve to valve and also in the course of time in one and the same valve.

The only two workers who appear to have attempted to enter the true retarding-field region in connection with fluctuation measurements are Williams (1936) and North (1940). The former operated with currents above 25 μA. while the theoretical value (for $v_a = 0$) of I_c was $\sim$50 μA.; but in view of the facts just discussed, the actual value of I_c may well have been considerably lower, which would, of course, explain that the quantity RI had not yet reached constancy in North's experiments. Williams' fluctuation measurements were carried out under the assumption that the critical value for the current in his valve was $\sim 50\,\mu$A. An estimate, however, on the basis of his valve data gives a theoretical value (for $v_a=0$) for I_c of the order of $1\,\mu$A. Further, in Williams' work over the current range 50–$12\,\mu$A. the factor RI diminished steadily and the indicated temperature (see § 5) was very much too high, all of which confirms that the retarding regime was not in fact entered in his work.

§ 5. DETERMINATION OF CATHODE TEMPERATURES FROM RETARDING FIELD MEASUREMENTS

It follows from equation (4) that in the retarding-field region

$$\log I = \frac{eV}{kT} + \text{const.} \qquad \ldots\ldots(38)$$

It appears, therefore, that by plotting $\log I$ against V and measuring the slope of this curve, the cathode temperature T can be determined. Work of this kind has been undertaken in the past by Möller and Detels (1926), Heinze and Hass (1938), Demski (1929), and others. The results of these experiments agree in general with temperature estimates based on other measurements. One general disadvantage of this method, however, is the use of a logarithmic plot which can easily mask a slow "power" variation of a variable, and therefore lead to considerable errors in the measurement of cathode temperatures.

On the other hand, it follows from equation (5) that the cathode temperature can be obtained from a plot of the differential resistence R against $1/I$, or of (RI) against I:

$$T = \frac{e}{k}(RI), \qquad \ldots\ldots(39)$$

provided that I is smaller than the critical current I_c. Evidently this method does not suffer from the source of error mentioned and should therefore give much more accurate results.

In order to demonstrate the application of this method, a scale of indicated temperatures derived from equation (39) is appended to figures 4, 5 and 6. The indicated temperatures are 1060° K. in the first case and 1150° K. in the second case, which agrees well with the standard specification of the valves and with the temperature values calculated from a formula by Widdell (1940) based on the radiation law.

The indicated temperatures in the experiments illustrated by figure 6 are 1000°, 880°, 710° K. respectively, which are in the ratios 10 : 8·8 : 7·1. According to Widdell's formula, the temperatures should be approximately in the ratios $W^{\frac{1}{4}}$ where W is the filament power (which is given at the bottom of figure 6). This leads to the ratios 10 : 8·7 : 7·3, showing excellent agreement and further illustrating the value of the method.

REFERENCES

DEMSKI, A., 1929. *Phys. Z.*, **30**, 291.

FRY, T. C., 1921. *Phys. Rev.*, **17**, 441.

FÜRTH, R. and MACDONALD, D. K. C., 1946. *Nature, Lond.*, **157**, 841.

HEINZE, W. and HASS, W., 1938. *Z. Tech. Phys.*, **19**, 166.

JOHNSON, J. B., 1925. *Phys. Rev.*, **26**, 71.

Langmuir, I., 1923. *Phys. Rev.*, **21**, 419.
v. Laue, M., 1918. *Jahrb. Rad. Elekt.*, **15**, 205.
Llewellyn, F. B., 1930. *Proc. Inst. Rad. Engrs.*, **18**, 243.
Möller, H. G. and Detels, F., 1926. *Jahrb. draht. Teleg. Teleph.*, **27**, 74.
North, D. O., Harris, W. A. and Thompson, B. J., 1940. *R.C.A. Rev.*, **4**; 1941. *Ibid.*, **5**; 1942. *Ibid.*, **6**.
Schottky, W., 1918. *Ann. Phys., Lpz.*, **57**, 541.
Schottky, W., 1926. *Phys. Rev.*, **28**, 74.
Schottky, W. and Spenke, E., 1937. *Wiss. Veröff. aus Siemens-Werken*, **16**, 1.
Wheatcroft, E. L. E., 1940. *J. Inst. Elect. Engrs.*, 473.
Widdell, 1940. *R.C.A. Rev.*, **5**, 106.
Williams, F. C., 1936. *J. Inst. Elect. Engrs.*, **78**, 326.
Williams, N. H. and Huxford, W. S., 1929. *Phys. Rev.*, **33**, 773.
Williams, N. H. and Vincent, H. B., 1926. *Phys. Rev.*, **28**, 1250.

STATISTICAL ANALYSIS OF SPONTANEOUS ELECTRICAL FLUCTUATIONS

By R. FÜRTH* and D. K. C. MacDONALD†,
* Department of Mathematical Physics, Edinburgh University;
now at Birkbeck College, London University
† Military College of Science; now at the Clarendon Laboratory,
Oxford University

MS. received 9 *October* 1946; *read* 21 *February* 1947

ABSTRACT. Electrical fluctuations, generated either as " shot-effect " in a saturated diode or as " thermal fluctuations " in a tuned circuit, were produced in a receiver of high natural frequency (~100 kc./s.-1 Mc./s.) and narrow band-width (~1-6 kc./s.), and (after suitable amplification) photographically recorded by means of a cathode-ray oscillograph operating on the single-stroke system. In these circumstances the fluctuations have the character of rapid oscillations (with the natural frequency of the receiver) whose amplitude R varies slowly and irregularly in time. The study of the statistical properties of this time variation of R, as represented by the envelope of the fluctuation record, was the subject of the present investigation. In particular, the distribution function of R within a statistically stationary series of observations and the correlation between values of R separated by a finite time interval were thoroughly investigated and, in general, found to be in good agreement with the statistical theory of the phenomenon in question.

§1. INTRODUCTION

A MECHANICAL system in static equilibrium will, under close observation, always exhibit irregular fluctuations about the equilibrium position which are usually referred to as *Brownian* motion. The statistical properties of this phenomenon have been thoroughly investigated in the past.‡ They can be grouped into two classes which we, using Smoluchowski's terminology, shortly denote by the words *magnitude* and *rate* of the Brownian motion. To the first class belongs the statistical distribution of the displacements from equilibrium as observed during a long interval of time, and the averages derived from this distri-

---

‡ See those references marked with an asterisk on pp. 402 and 403.

bution. To the second class belongs the probability function for observing a transition from one position to another during a fixed time interval and the correlation averages derived therefrom. The magnitude of the Brownian motion depends solely on the inertia of the system and the forces acting upon it, and on the temperature, but not on friction and other dissipative agencies. The rate of the Brownian motion, on the other hand, is essentially governed by these latter properties, which is especially evident in the ordinary Brownian motion of free particles in colloidal solutions.

A number of experimental investigations on various systems of the type mentioned has been carried out in which the Brownian motion of the system was continuously recorded. Statistical analysis of such records have completely confirmed every aspect of the theory. The experiments of E. Kappler (1932) on the Brownian motion of a torsion balance are the most complete of this kind; here the rate of the Brownian motion could be changed independently of its magnitude by altering the damping of the system by the variation of the pressure of the surrounding air.

The analogous phenomena in electrical circuits, which in certain limiting cases manifest themselves either as " thermal fluctuation of current " or as " shot-effect ", were first predicted in a classical paper on Brownian movement by A. Einstein (1906), and have more recently been the subject of many theoretical and experimental investigations (see e.g. Moullin, 1938). It could be shown in particular that the magnitude of the current fluctuations as expressed by the mean square of the deviations of the current from its average was in accord with theory, but no work aiming at a complete experimental statistical analysis of electrical fluctuations with regard to magnitude and rate has been published so far. This is primarily due to one difficulty inherent in this phenomenon. Whereas the Brownian motion of a mechanical system can be comparatively easily observed by optical methods (e.g. by light-beam amplification) without affecting the motion to any measurable degree, the electrical fluctuations can only be observed by means of some electrical measuring instrument which must be connected to the circuit under observation. This, however, not only alters the electrical characteristics of the circuit but also introduces a distortion of the original fluctuations owing to the mechanical properties of the measuring instrument; moreover, the mechanical Brownian motion of the measuring instrument cannot be separated from the electrical fluctuations which it should measure. These problems have been discussed, mainly by Ornstein (1927, 1938) and his collaborators (Ornstein, Burger and Taylor, 1927), in a series of brilliant theoretical and experimental investigations.

More recently, however, electronic methods of amplification and recording by means of cathode-ray oscillographs have been developed to such a perfection that electrical fluctuations can now be recorded in such a way that only the electrical properties of the circuit (including the amplifiers) are involved. Thus the way to a thorough statistical analysis of electrical fluctuations in complete analogy to the above-mentioned work on mechanical Brownian movement is opened, and some first steps in this direction have been taken by the present authors. Since on the other hand the theory of the statistics of electrical fluctuations covering both the magnitude and the speed have been developed to a high degree of perfection, for example in papers by Wang and Uhlenbeck (1945) and Rice (1945), a comparison between experiment and theory could be carried out in order to verify the underlying general principles of the theory.

In what follows, this work, on which we published a preliminary note (1946), will be described in some detail. §2 gives a description of the experimental method and the procedure for analysing the records; §3 contains the investigation on the magnitude of the fluctuations; §4 some theoretical considerations on the problem of the rate of the fluctuations; and §5 the comparison between the experiments and the results of these considerations.

§2. EXPERIMENTAL METHOD

In the experiments of Kappler (1932), the Brownian motion of an oscillatory system could be directly observed and recorded, as the periodic time of the torsion balance was of the order of magnitude of 30 sec. For low damping, the records have the character of harmonic oscillations, at the natural frequency of the oscillating system, whose amplitude varies irregularly in time, the speed of this variation depending on the amount of damping. The main subject of interest in this case is, therefore, not the shape of the actual oscillation record, but that of the "envelope" of the fluctuation curve which represents the variation of amplitude in time.

In order to be able to observe true electrical fluctuations it is necessary to avoid the interference of other irregular disturbances of a different nature. The most significant of these is the so-called *Flicker effect* (Johnson, 1925) which arises from some relatively gross variation of the emissive properties of valve emitters. Similar effects also occur for instance in semi-conductors. As these disturbances, which have nothing to do with the phenomenon in question, are known to vary at a comparatively slow rate, they can be effectively excluded by using oscillating circuits of high natural frequency and low damping for the generation of the electrical fluctuations. Under such circumstances, the records of the fluctuations will evidently again have the character of harmonic oscillations, the amplitude of which will vary irregularly in time at a rate determined by the overall "band-width" of the circuit plus amplifiers and recorder. It will therefore suffice to record the envelope of the fluctuation curve for fluctuations generated either by shot-effect in a diode or by thermal fluctuation in a metallic circuit, and for various band-widths.

It was decided to use a standard Marconi Communication Receiver CR/100/2 in connection with a Cossor double-beam cathode-ray oscillograph as a recording instrument. The primary fluctuations were generated in the input circuit of this receiver at various frequencies between about 100 kc./s. and 2 Mc./s. The "effective" natural frequency, however, was the customary intermediate frequency of the receiver—465 kc./s.—which is high enough to avoid the aforesaid disturbing effects. The amplification per stage of the receiver for such a frequency is also sufficiently high for the fluctuations introduced by the amplifier valves not to affect the records seriously. The band-widths provided in this receiver are normally 6 kc./s., 3 kc./s., 1·2 kc./s. and 0·3 kc./s., and were found to be very suitable for the purpose of this investigation. It was first attempted to record the detected output of the receiver which should essentially represent the amplitude variations in time. It was found, however, that considerable detail was lost in this process and, therefore, no detector was used in the final experiments, although

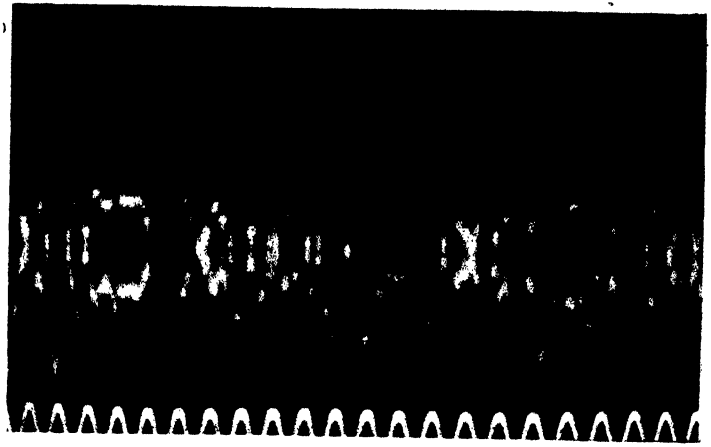


Figure 1. Photographic record of shot-fluctuation. Dominant frequency 100 kc./s., nominal band-width 6 kc./s., timing wave 1 kc./s.

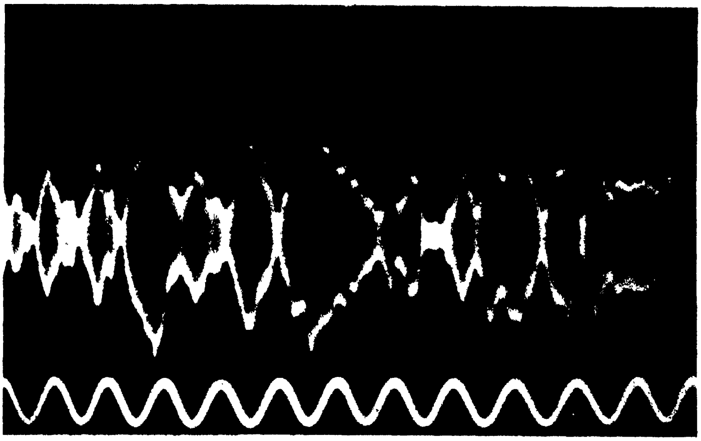


Figure 2. Photographic record of thermal fluctuation. Dominant frequency 130 kc./s., nominal band-width 1·2 kc./s., timing wave 500 c./s.

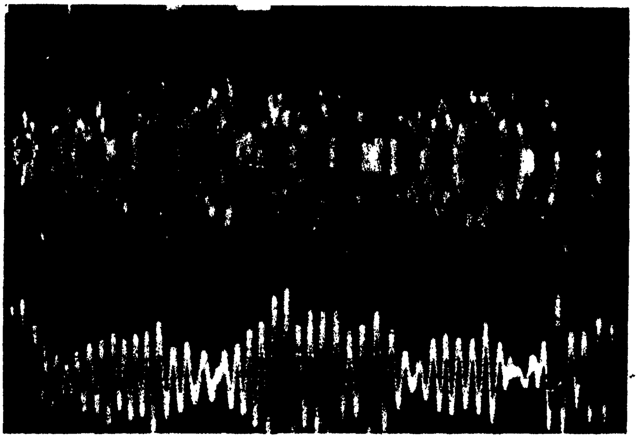


Figure 3. Upper part : record of spontaneous electrical fluctuation. Lower part : the same fluctuation, but rectified and passed through low-frequency filter with narrow response.

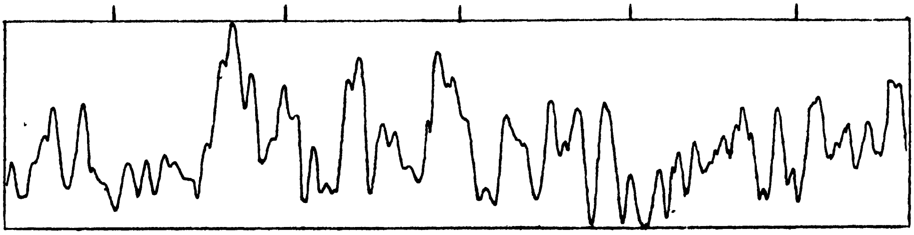


Figure 4. Traced envelope of " shot " fluctuation record. Dominant frequency 850 kc./s., nominal band-width 6 kc./s., timing marks 1 m.s.

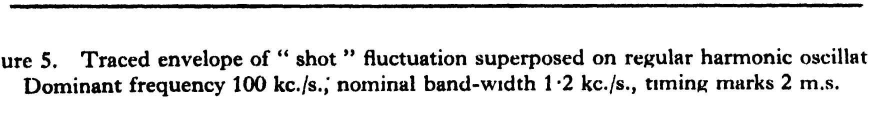


Figure 5. Traced envelope of " shot " fluctuation superposed on regular harmonic oscillation Dominant frequency 100 kc./s., nominal band-width 1·2 kc./s., timing marks 2 m.s.

it is hoped at some later date to examine more exhaustively the design of a rectifier for such work. Owing to the rapidity of the movement of the spot on the oscillograph screen, apart from the neighbourhood of the extreme points of the oscillation, the records taken in this way show practically only the envelope curve (see figures 1 and 2).

Some experiments were undertaken, however, in order to study the character of the fluctuations obtained by letting the rectified original fluctuations pass through a low-frequency filter of limited response. A diode detector was used and the filter employed had a practically "Gaussian" response curve centred about 1000 c./s. Figure 3 gives an example of records obtained in this way together with the record of the original unrectified fluctuations. As might be expected, the phenomenon has now the character of a harmonic oscillation of 1000 c./s. frequency with amplitude slowly varying at a rate determined by the damping of the filter, and the records have a striking analogy with those obtained by Kappler for very light damping of his system.

The photographs were taken by the "single-stroke" system, incorporating a modification of the "trigger" facility already provided in the Cossor oscillograph. This method is preferable to the method of recording on a continuously moving film (as used by Kappler in his investigations) because no special camera has to be used and because the afterglow of the oscillograph screen has no ill effect in the former method. It was further found that sufficient statistical material could be obtained from one single "frame". The second beam of the oscillograph was, as a rule, used for producing a timing wave on the records, which was achieved by feeding the output from a beat-frequency oscillator on to the corresponding deflecting plates. Frequencies of 250, 500, 1000 and 2000 c./s. were used, and the timing waves are exhibited at the bottom of the records (see, for example, figures 1 and 2). In some cases the two beams were used for the simultaneous recording of two different fluctuation phenomena (see, for example, figure 3).

Photographs were taken using both a Leica and a Cossor camera; the exposure time was effectively determined by the duration of the "single-stroke" time-base and varied between about 1/10 sec. and 5/1000 sec., according to the rate of the fluctuation as governed by the band-width and the detail required. To carry out detailed analysis of the records, the original negatives were placed in a standard Leitz enlarger, and the envelopes of the curves traced on drawing paper in pencil and later traced over in Indian ink. Examples of such records are given in figures 4 and 5.

§ 3. THE MAGNITUDE OF THE FLUCTUATIONS

The first analysis was conducted to examine the distribution of R within the envelope curve $R(t)$ over a sufficiently large observation time. It was shown by Rice (1944, 1945) that the probability $p(R)dR$ for values of R within the interval $(R, R+dR)$ is given by

$$p(R) = \frac{R}{\psi} e^{-R^2/2\psi} \left(\int_0^\infty p(R)dR = 1 \right), \qquad \ldots\ldots(1)$$

where the constant ψ is defined by

$$\psi = \int_0^\infty W(f)df, \qquad \ldots\ldots(2)$$

$W(f)$ being the density in the "power spectrum" of the current fluctuations, and f the frequency.*

In order to check this theoretical prediction, a method was adopted which has been frequently used for similar purposes in the past (for example, Fürth, 1917; Kappler, 1932). Horizontal lines were drawn on the records at uniform vertical intervals (as seen in figures 4 and 5) and the number of intersections $q(R)$ with the trace noted. An example of a distribution obtained in this way is shown in figure 6.

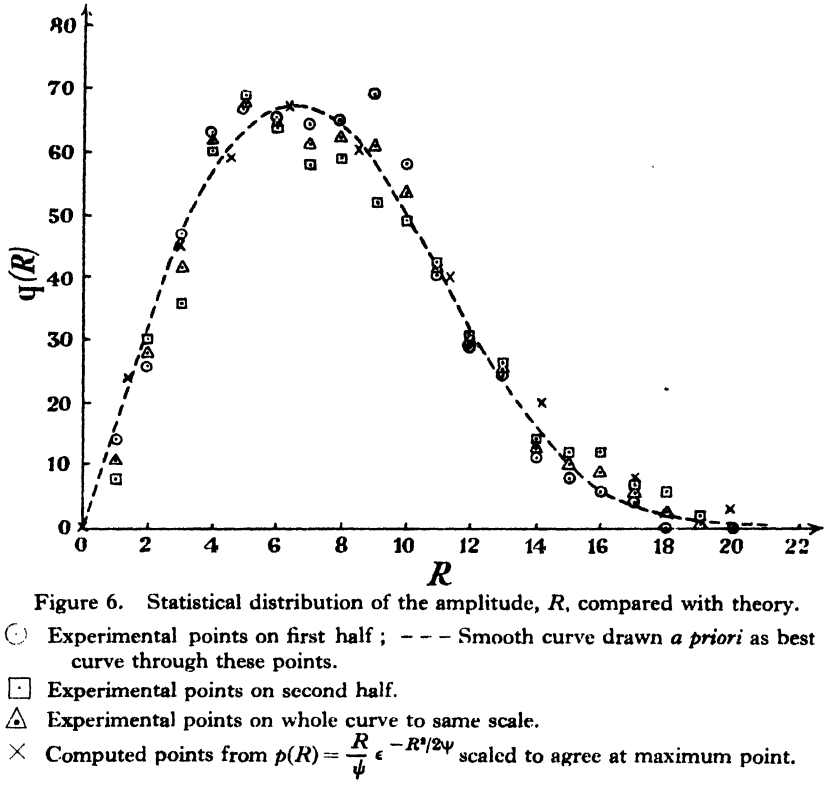


Figure 6. Statistical distribution of the amplitude, R, compared with theory.

⊙ Experimental points on first half ; – – – Smooth curve drawn *a priori* as best curve through these points.

⊡ Experimental points on second half.

△ Experimental points on whole curve to same scale.

× Computed points from $p(R) = \frac{R}{\psi}\epsilon^{-R^2/2\psi}$ scaled to agree at maximum point.

It can be shown that the distribution $q(R)$ is, apart from a constant factor, identical with the distribution $p(R)$ provided that the slope dR/dt of the curve $R(t)$ is completely uncorrelated with R everywhere†; thus the procedure of identifying

* Formula (1) can be derived from the general distribution formula for the fluctuation current I, that is, the probability $p(I)dI$ that $I(t)$ lies between I and $I+dI$ which, according to Rice, is

$$p(I)dI = \frac{1}{\sqrt{2\pi\psi}}e^{-I^2/2\psi}\,dI.$$

From this it follows that the mean square of the current is equal to $\overline{I^2} = \psi$. In the particular case considered here, where the band-width of the system is small compared with the mid-frequency, the current fluctuation can be regarded as the superposition of sine and cosine oscillations (at the dominant frequency of the system), whose amplitudes vary irregularly and independently in time according to the above formula.

† Call δt_1, δt_2, ... the time intervals during which $R(t)$ is found in the narrow range between R and $R+\delta R$ within the total observation time T. Then by definition

$$p(R)\delta R = (\delta t_1 + \delta t_2 + \ldots .)/T = \left(\frac{\delta t_1}{\delta R} + \frac{\delta t_2}{\delta R} + \ldots .\right)\frac{\delta R}{T},$$

from which follows

$$p(R) = \frac{1}{T}\left(\overline{\frac{dt}{dR}}\right)q(R).$$

If, now, dR/dt is uncorrelated with R, the quantity $\overline{(dt/dR)}$ is independent of R and thus $p(R)$ is proportional to $q(R)$, as stated in the text.

$q(R)$ with $p(R)$ is justified under this condition but completely unjustified, e.g., for a periodic curve $R(t)$. In order to make sure about this point in the present problem a " detailed " analysis was also carried out in some cases by subdividing the t-axis into narrow intervals (as seen in figure 4) and determining the distribution of the corresponding values of R. The result was essentially the same as that obtained by the first method, which was accordingly generally adopted in view of its greater simplicity.

Before going further, another necessary check had to be made to ascertain that the fluctuation was statistically stationary over the whole observation time and that a possible burst of local interference had not vitiated the record. To this end the records were frequently divided into two halves and the distribution of each half studied separately to see whether they fitted to each other as demonstrated by figure 8. Any record which was found not to satisfy this test was discarded.

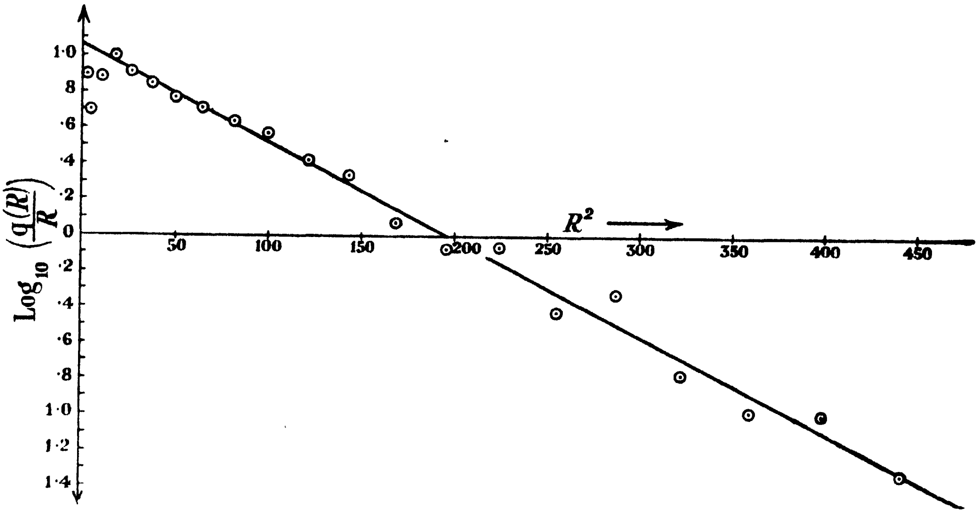


Figure 7. Logarithmic representation of the statistical distribution of amplitudes R.

The simplest way to prove whether the observed distribution agrees with the theoretical formula (1) is to draw a smooth curve through the observation points and fit the available constants so that the observed maximum of this curve coincides with the maximum of the function (1). This procedure is illustrated in figure 6, and one sees that there is no systematic deviation of the observed from the theoretical points. A more stringent test can be applied by plotting $\log[p(R)/R]$ against R^2, which, according to (1), should give points lying on a straight line with slope $-1/2\psi$. An example of such a plot is presented in figure 7, which shows that this is indeed the case, leading to the value $\psi=40$ for this particular set of observations.

The average moments and the most probable value R_m of R, i.e. that value of R for which $p(R)$ is a maximum, can be easily derived from formula (1); one has, for instance,

$$\left.\begin{array}{ll}\bar{R}=\sqrt{\pi\psi/2}, & \overline{R^3}=3\psi\sqrt{\pi\psi/2},\\ \overline{R^2}=2\psi, & R_m=\sqrt{\psi},\end{array}\right\} \qquad \ldots\ldots(3)$$

whence

$$\frac{(\bar{R})^2}{\overline{R^2}}=\frac{\pi}{\psi}, \qquad \frac{\overline{R^3}}{\bar{R}\,.\,\overline{R^2}}=\frac{3}{2}, \qquad \frac{R_m^2}{\overline{R^2}}=\frac{1}{2}. \qquad \ldots\ldots(4)$$

By direct computation from the observed values of R, the following "experimental" values were obtained for the above-mentioned set of observations :

$$\overline{R^2}\,(\text{exp}) = 81, \qquad R_m\,(\text{exp}) \sim 7,$$

which compare well with the theoretical values derived from (3):

$$\overline{R^2}\,(\text{theor}) = 80, \qquad R_m\,(\text{theor}) = 6{\cdot}3.$$

Another observation series yielded the following experimental values for the ratios of the moments:

$$(\bar{R})^2/\overline{R^2}\,(\text{exp}) = 0{\cdot}75, \qquad \overline{R^3}/\bar{R}\,.\,\overline{R^2}\,(\text{exp}) = 1{\cdot}7,$$

which again agree well with the theoretical values 0·785 and 1·5.

A further interesting problem which belongs to the same class of magnitude fluctuation phenomena is the distribution of the frequencies with which maxima of different heights occur in the envelope curve during a sufficiently long observation time. This problem has also been discussed by Rice (1944, 1945), and his analysis shows that this distribution function is solely determined by the electrical properties of the network concerned, as was to be expected. The experimental data for the derivation of the distribution function of the maxima were therefore collected for a number of "frames" and a typical example is shown in figure 8. The spread of the observational points is naturally very considerable owing to the limited number of maxima within a frame. No detailed computation was therefore carried out, but the general shape of the smooth curve drawn (rather arbitrarily) through the observational points agrees with the ones derived by Rice from his theory. Much longer series of observations would have to be carried out were a detailed check of the theory to be attempted.

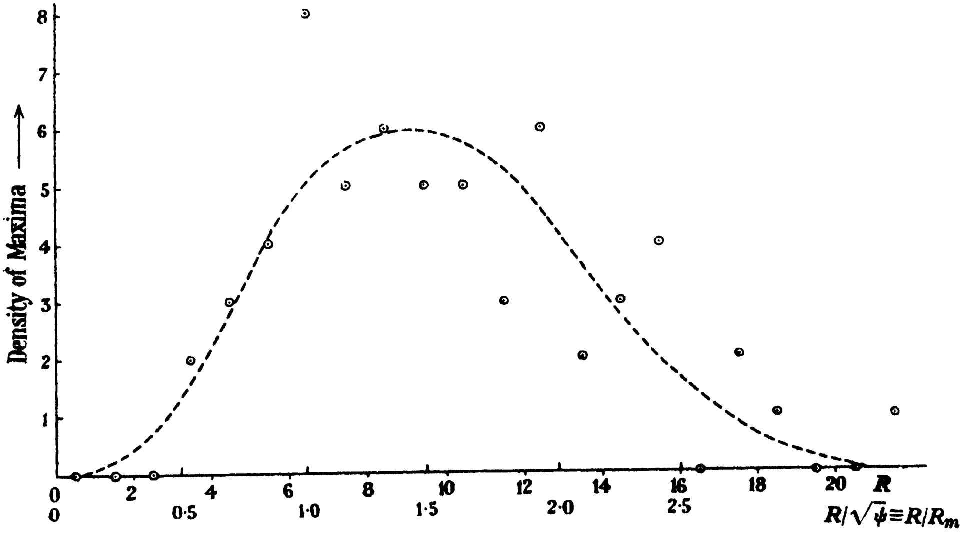


Figure 8. Statistical distribution of the heights of the maxima in the envelope records $R(t)$.

○ Experimental points.

- - - - Smooth curve through experimental points (extremely small maxima are omitted).

A special series of experiments was made in order to study the effect of the superposition of a regular harmonic oscillation on the irregular fluctuations.

For this purpose a regular "signal" (from a signal generator) was injected at the receiver input (tuned to the same frequency) and the fluctuations were recorded in the usual way.

The theory of this phenomenon is also presented in Rice's papers. He obtains for the distribution function $p(R)$:

$$p(R)=\frac{R}{\psi}\exp\left[-\frac{R^2+S^2}{2\psi}\right].I_0\left(\frac{RS}{\psi}\right), \qquad \ldots\ldots(5)$$

where S is the amplitude of the signal as it would appear on the record when no fluctuations were present, and where I_0 is the Bessel function of zero order with imaginary argument. From (5) follows, for the mean square of R,

$$\overline{R^2}=S^2+2\psi \qquad \ldots\ldots(6)$$

as a direct consequence of the fact that the regular oscillations and the random oscillations are not correlated. When S becomes "large", one can replace $I_0(z)$ by its asymptotic expression:

$$\lim_{z\to\infty} I_0(z)\approx\frac{e^z}{\sqrt{2\pi z}} \qquad \ldots\ldots(7)$$

(apart from the very beginning of the distribution curve), and (5) goes over into

$$p(R)\sim\frac{1}{\psi}\sqrt{\frac{R}{2\pi S}}\exp\left[-\frac{(R-S)^2}{2\psi}\right], \qquad \ldots\ldots(8)$$

which shows that the fluctuation amplitude is almost normally distributed about the amplitude of the regular oscillation.

Figure 5 is a typical example of a record of $R(t)$ obtained in these experiments, and figure 9 shows the corresponding distribution function. The computation of the mean square of R gave $\overline{R^2}=239{\cdot}4$ and the abscissa of the maximum of the curve gives $S=14{\cdot}3$. Thus from (6), $2\psi=35{\cdot}4$ and $a=S/\sqrt{\psi}=3{\cdot}4$. Rice has computed the function $p(R)$ for various values of a from (5) and the result is

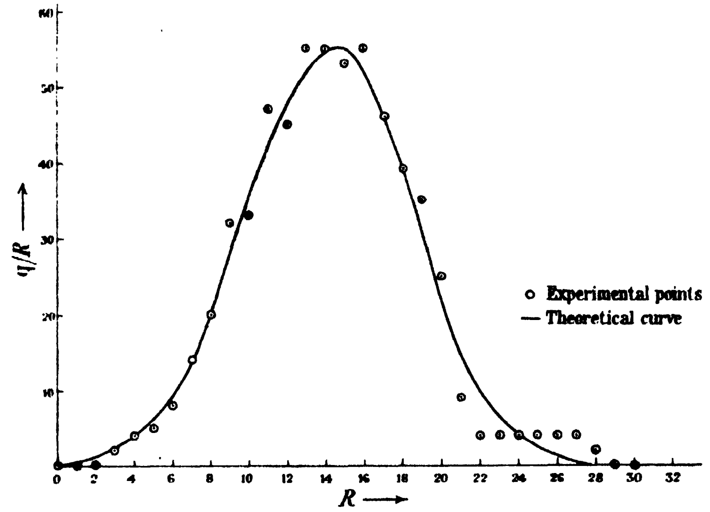


Figure 9. Statistical distribution of the amplitudes R of a fluctuation superposed on a regular harmonic oscillation, compared with theory.

presented in his paper in a set of curves. As these only vary slowly in shape as the parameter a varies, the curve relating to $a=3$ (which is nearest to the actual value 3·4) was taken and reduced to scale so as to make the maxima of the experimental and the theoretical curves coincide. This theoretical curve is shown in figure 11 and is seen to fit the observational points very satisfactorily.

§4. SOME THEORETICAL CONSIDERATIONS REGARDING THE FLUCTUATION RATE

We now discuss the second class of fluctuation phenomena which, as indicated in § 1, are connected with the rate at which the fluctuations occur or, in other words, with the degree of correlation between fluctuations separated by a finite time interval. If it is assumed that the primary effects which give rise to the fluctuations are completely uncorrelated, this correlation will be entirely due to the electrical characteristics of the network used. If, on the other hand, these primary effects are already correlated to some degree, the observed rate would also be affected by this primary correlation.

The theory of the rate phenomena under the assumption of complete randomness of the primary effects has been developed by Rice in particular. R. E. Burgess has also contributed to this field in unpublished work. Some further aspects of the theory have been developed by the present authors. The object of the analysis of the fluctuations with respect to rate was to check these theoretical relations and to see whether there are any indications that the correlation of the primary effects might have to be taken into account.

The basic formula to start from is that for the probability $p(R_1, R_2)dR_1dR_2$, of observing a value between R_1 and R_1+dR_1 at time t and a value between R_2 and R_2+dR_2 at time $t+\tau$ within a statistically stationary series of observations. This is, according to Rice,

$$p(R_1R_2)=\frac{R_1R_2}{A}\cdot I_0\left[\frac{R_1R_2}{A}(\mu^2+\lambda^2)^{\frac{1}{2}}\right]\exp\left[-\frac{\psi}{2A}(R_1{}^2+R_2{}^2)\right], \quad \ldots\ldots(9)$$

where ψ is the quantity defined by (2) and where μ, λ are the sine and cosine Fourier transforms respectively of the power-spectrum density function $W(f)$:

$$\left.\begin{aligned}\mu(\tau)&=\int_0^\infty W(f)\cos 2\pi(f-f_0)\tau\,.\,df,\\ \lambda(\tau)&=\int_0^\infty W(f)\sin 2\pi(f-f_0)\tau\,.\,df,\end{aligned}\right\} \quad \ldots\ldots(10)$$

f_0 being the dominant frequency of the network. A is an abbreviation:

$$A=\psi^2-\mu^2-\lambda^2, \quad \ldots\ldots(11)$$

and I_0 is again the Bessel function of zero order with imaginary argument.

Rice calculated $p(R_1, R_2)$ in particular for the case of an ideal band-pass filter with a response curve of rectangular shape. In the present case, where the response curve of the receiver used was very approximately "Gaussian" in shape, we can set

$$W(f)=\frac{\psi}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(f-f_0)^2}{2\sigma^2}\right], \quad \ldots\ldots(12)$$

which satisfies (2) when $\sigma \ll f_0$. Under the same condition one obtains easily from (10), (11), (12):

$$\mu=\psi z, \quad \lambda=0, \quad A=\psi^2(1-z^2), \quad \text{where } z=e^{-2\sigma^2\pi^2\tau^2}. \quad \ldots\ldots(13)$$

Introducing the quantity

$$v_\tau = R_2 - R_1, \qquad \ldots\ldots(14)$$

we can now define the probability $P(v_\tau)dv_\tau$ for observing a fixed difference between v_τ and $v_\tau + dv_\tau$ of R over a given time interval τ, irrespective of R_1 within the series of observations:

$$P(v_\tau) = \int_0^\infty p(R_1, R_1 + v_\tau)dR_1 \qquad (v_\tau > 0) \qquad \ldots\ldots(15)$$

which, of course, will be independent of time. As (9) is symmetrical in R_1, R_2, obviously $P(-v_\tau) = P(v_\tau)$.

The average value Δ_τ of $|v_\tau|$ is defined by

$$\Delta_\tau = \overline{|v_\tau|} = 2\int_0^\infty v_\tau \,.\, P(v_\tau)dv_\tau, \qquad \ldots\ldots(16)$$

for which in analogous problems of Brownian motion the term "after-effect" has been introduced by Smoluchowski (1923). The quantity

$$\epsilon = \lim_{\tau=0} (\Delta_\tau/\tau)$$

is a convenient measure for the overall speed of the fluctuations. It may be remarked, however, that the fluctuations will still take place at a finite rate even if this quantity is zero or infinite.

We first consider the case of τ very small. Under this condition we obtain from (13) approximately

$$\mu = \psi(1 - 2\pi^2\sigma^2\tau^2), \qquad A = 4\pi^2\sigma^2\tau^2\psi^2. \qquad \ldots\ldots(17)$$

As now the argument of I_0 in (9) becomes very large except for small values of R_1R_2, we can again use the asymptotic expression (7); hence

$$\lim_{\tau\to 0} p(R_1R_2) = \frac{\sqrt{R_1R_2}}{(2\pi\psi)^{3/2}\sigma\tau} \exp\left[-\frac{R_1R_2}{2\psi} - \frac{(R_1 - R_2)^2}{8\pi^2\psi\sigma^2\tau^2}\right]. \qquad \ldots\ldots(18)$$

From (14), (15) and (18) we obtain by simple calculation

$$\lim_{\tau=0} P(v_\tau)dv_\tau = \frac{1}{\sqrt{\pi}} e^{-x^2} dx \quad \left(\int_{-\infty}^{+\infty} P(v_\tau)dv_\tau = 1\right), \qquad \ldots\ldots(19)$$

where

$$x = \frac{v_\tau}{2\pi\sigma\tau\sqrt{2\psi}}, \qquad \ldots\ldots(20)$$

which shows that the quantity v_τ is normally distributed about zero, as was to be expected.

From (16) we now get

$$\lim_{\tau\to 0} \Delta_\tau = 2\sqrt{2\pi\psi} \,.\, \sigma\tau, \qquad \ldots\ldots(21)$$

which shows that the "speed" ϵ of the fluctuations is finite and equal to

$$\epsilon = \sqrt{2\pi\psi} \,.\, \sigma.$$

We now turn to the opposite limiting case of τ very large. Here we have from (13) $\mu = 0$, $A = \psi^2$, and since $I_0(0) = 1$, formula (9) becomes

$$\lim_{\tau\to\infty} p(R_1R_2) = \frac{R_1R_2}{\psi^2} e^{-\frac{R_1^2 + R_2}{2\psi}}, \qquad \ldots\ldots(22)$$

which also follows immediately from (1), as in this case there is evidently no correlation between the fluctuations of R_1 and R_2.

Similarly, as above, we obtain now

$$\lim_{\tau\to\infty} P(v_\tau)\,dv_\tau = e^{-y^2}\{ye^{-y^2}+\sqrt{\pi}\,(\tfrac{1}{2}-y^2)(1-\operatorname{erf}y)\}\,dy, \qquad ......(23)$$

where

$$y=\frac{|v_\tau|}{2\sqrt{\psi}} \qquad ......(24)$$

and

$$\operatorname{erf}y\equiv\frac{2}{\sqrt{\pi}}\int_0^y e^{-u^2}\,du \qquad ......(25)$$

is the well known error function.

In the general case* of an arbitrary τ it is advantageous to introduce instead of R_1 the new variable

$$t=\frac{R_1(R_1+v_\tau)}{\psi(1-z^2)} \qquad ......(26)$$

from (9), (13), (15), and one now obtains easily

$$P(v_\tau)=(1-z^2)\,.\exp\left[-\frac{v_\tau^{\,2}}{2\psi(1-z^2)}\right]\int_0^\infty \frac{te^{-t}I_0(zt)}{\sqrt{v_\tau^{\,2}+4\psi(1-z^2)t}}\,dt. \qquad ......(27)$$

Using the identity

$$\frac{1}{\sqrt{\pi}}\int_0^\infty \xi^{-\frac{1}{2}}e^{-\alpha\xi}\,d\xi\equiv\frac{1}{\sqrt{\alpha}} \qquad ......(28)$$

we can transform (27) into

$$P(v_\tau)=\frac{\sqrt{1-z^2}}{2\sqrt{\pi\psi}}\,.\exp\left[-\frac{v_\tau^{\,2}}{2\psi(1-z^2)}\right].\int_0^\infty \xi^{-\frac{1}{2}}e^{-\frac{v_\tau^2\xi}{4\psi(1-z^2)}}\,d\xi\int_0^\infty te^{-t(1+\xi)}I_0(zt)\,dt.$$

$$......(29$$

Now it may be shown by Hankel's formula (Whittaker and Watson, 1935) that

$$\int_0^\infty t\,.\,e^{-t(1+\xi)}\,I_0(zt)\,dt=\frac{1+\xi}{\xi^{\frac{1}{2}}[(1+\xi)^2-z^2]^{3/2}}\,. \qquad ......(30)$$

Thus finally,

$$P(v_\tau)=\frac{\sqrt{1-z^2}}{2\sqrt{\pi\psi}}\exp\left[-\frac{v_\tau^{\,2}}{2\psi(1-z^2)}\right]\int_0^\infty \frac{1+\xi}{\xi^{\frac{1}{2}}[(1+\xi)^2-z^2]^{3/2}}\,e^{\frac{v_\tau^2\xi}{4\psi(1-z^2)}}\,d\xi.$$

$$......(31)$$

From (16) and (31) we get for the after-effect for arbitrary τ

$$\Delta_\tau=2\sqrt{\frac{\psi(1-z^2)^3}{\pi}}\int_0^\infty\frac{1+\xi}{\xi^{\frac{1}{2}}(2+\xi)[(1+\xi)^2-z^2]^{3/2}}\,d\xi \quad (z=e^{-2\pi^2\sigma^2\tau^2}). \qquad ......(32)$$

In particular for $\tau\to\infty$ (the case considered above) one has

$$\Delta_\infty=2\sqrt{\frac{\psi}{\pi}}\int_0^\infty\frac{d\xi}{\xi^{\frac{1}{2}}(1+\xi)^2(2+\xi)}=(\sqrt{2}-1)\sqrt{\pi\psi}=0{\cdot}73\sqrt{\psi}. \qquad ......(33)$$

This differs from $\bar{R}$ (formula (3)) only by a numerical factor because of the complete lack of correlation for long time intervals.

For values of τ not too small the quantity z is small and the integral in (32) can be expanded into a power series with respect to z:

$$\Delta_\tau=(\sqrt{2}-1)\sqrt{\pi\psi(1-z^2)^3}\,\{1+1{\cdot}035z^2+1{\cdot}046z^4+1{\cdot}046z^6+\ldots\},$$

---

* We have to thank Dr. A. Erdélyi for his valuable advice on the solution of this problem.

which is approximately equal to

$$\Delta_\tau=(\sqrt{2}-1)\sqrt{\pi\psi(1-z^2)}=(\sqrt{2}-1)\sqrt{\pi\psi}\,\sqrt{1-e^{-4\pi^2\sigma^2\tau^2}}. \quad \ldots\ldots(34)$$

This, of course, goes over into (33) for $\tau=\infty$. For small values of τ, on the other hand, the expression on the right-hand side of (34) becomes

$$(\sqrt{2}-1)\sqrt{\pi\psi}\,.\,2\pi\sigma\tau=0{\cdot}84\,.\,2\sqrt{2\pi\psi}\,.\,\sigma\tau,$$

which differs from the correct expression (21) by the numerical factor 0·84. Thus it is seen that formula (34) represents the function Δ_τ with very good approximation over almost the whole range of τ from 0 to ∞.

§ 5. COMPARISON BETWEEN THE THEORY OF THE FLUCTUATION RATE AND EXPERIMENT

The theoretical results of § 4 were put to the test by subdividing the *t*-axes of the fluctuation records into equal time intervals of length τ and reading off the differences v_τ of the successive ordinates. Thus the frequency distribution of the different values of v_τ within a record and the average Δ_τ can be obtained and compared with the theoretical formulae.

First the limiting case of small τ was studied on a number of fluctuation records, using time intervals corresponding to a distance of 1 mm. on the records, which were just large enough to make the measurements of the ordinates reasonably accurate. The distribution of v_τ for two records, one of "shot" fluctuations, the other of "thermal" fluctuations, is displayed in figure 10, together with the theoretical Gauss distribution curve (19) drawn in such a way as to give the best fit of the observed points. This shows that formula (19) is indeed satisfied. Moreover, the value of the constant $C=v_\tau/x$ can now be determined for each record and compared with its theoretical value $2\pi\sigma\tau\sqrt{2\psi}$ according to (20). τ is readily determined from the timing wave on the record, and $\psi=\overline{R^2}/2$ (see formula (3)) is found by the method described under § 3. σ can be determined from the frequency-response curve of the receiver, observed under the same operating condition under which the fluctuation record is obtained, and by comparing this curve with formula (12) by numerical integration. The results for three records, including the two mentioned above, are collected in table 1 and show a very satisfactory agreement between the theoretical and experimental values of C.

Table 1

Record	$\tau\times10^5$ sec.	$\sqrt{\overline{R^2}}$	$\sigma\times10^{-3}$ sec.$^{-1}$	C (theor.)	C (exp.)	$\dfrac{C\text{ (exp.)}}{C\text{ (theor.)}}$
Shot fluctuation (4—J)	2·2	9·0	1·95	2·42	2·77	1·14
Thermal fluctuation (7—J)	2·24	9·6	2·08	2·76	2·76	1·00
Thermal fluctuation (16—C)	4·5	9·4	1·19	3·14	3·26	1·04

We now turn to the experimental results for large intervals. To examine the frequency distribution of v_τ, the differences of the ordinates R at 30 mm. apart were employed for the first record mentioned above. As Δ_τ was found to reach a stationary value for $\tau > 10$ mm. say (see figure 12) in this case, the separation employed should correspond well to the theoretical case of $\tau \to \infty$. The observed distribution for this record is shown in figure 14. From the experimental value $\overline{R^2} = 81$ for this particular series we obtain $2\sqrt{\psi} = 12{\cdot}7$. Using this scale factor for the abscissae according to (24), the observed points should lie on the theoretical curve (23). This curve is also shown in figure 11, and the observed points are seen to fit the theoretical curve satisfactorily.

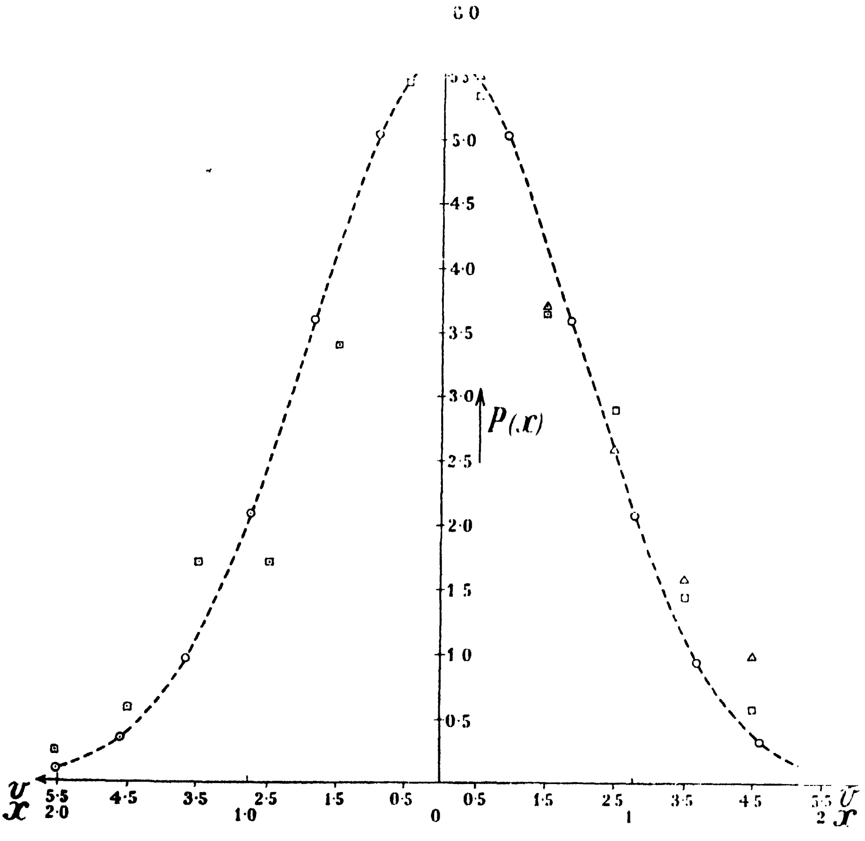


Figure 10. Statistical distribution of v_τ for small τ, compared with theory.

○ - - - } $P(x) = \frac{1}{\sqrt{\pi}} \epsilon^{-x^2}$.

□ Experimental points for shot fluctuation.

△ Experimental points for thermal fluctuation.

In addition, the limiting values of Δ_∞ (the average of $|v_\infty|$) were estimated from a number of records; at the same time the theoretical values of Δ_∞ were calculated from (33). These are exhibited in table 2, and again agreement between theory and experiment is established.

No check of the general distribution formula (31) was attempted, but formula (34) for Δ_τ was put to the test by computing the values of $\Delta_{\tau n}$ for the basic time interval τ (corresponding to 1 mm. spacing on the records) and its integer multiples

Table 2

Record	$\sqrt{\overline{R^2}}$	Δ_∞ (theor.)	Δ_∞ (exp.)	$\dfrac{\Delta_\infty\text{(exp.)}}{\Delta_\infty\text{(theor.)}}$
Shot fluctuation (4—J)	9·0	4·66	4·6	0·98
Shot fluctuation (6—J)	10·07	5·21	5·9	1·13
Thermal fluctuation (7—J)	9·06	4·98	4·8	0·96
Thermal fluctuation (7—J)	12·93	6·71	7·2	1·07

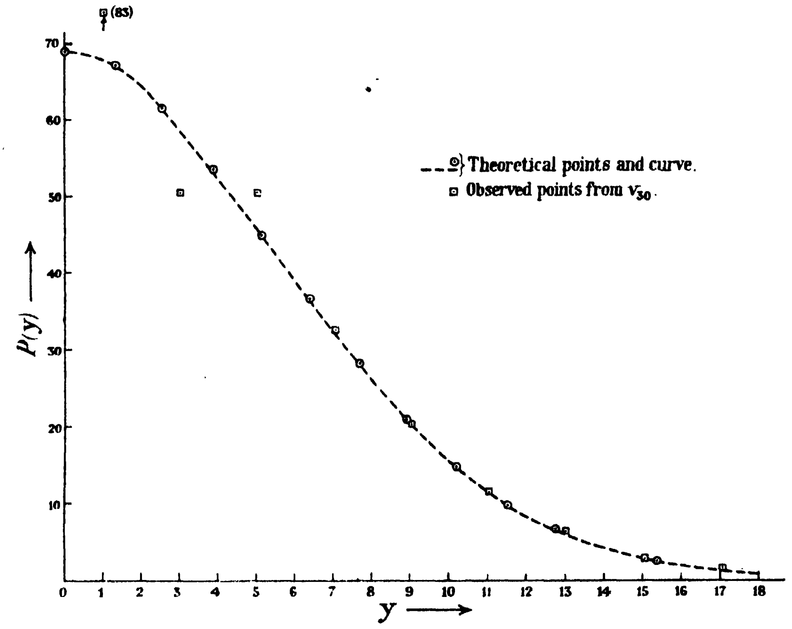


Figure 11. Statistical distribution of v_τ for large τ, compared with theory.

$n\tau$, again for a number of shot and thermal fluctuation records. In order to facilitate the comparison with theory the reduced after-effect factors

$$\delta_n = \frac{\Delta_{\tau n}}{\Delta_\infty} \qquad \ldots\ldots(35)$$

were introduced, which according to (33) and (34) should satisfy the formula

$$\delta_n = \sqrt{1 - e^{-4\pi^2\sigma^2\tau^2n^2}}. \qquad \ldots\ldots(36)$$

An example of the results is shown in figure 12. The broken line represents the function (36) for the values of the constants given in table 1. This function is inaccurate for small values of n, as explained in § 4. The beginning of the correct curve, on the other hand, can be calculated from (12). By joining this with the former curve, the correct theoretical curve is obtained drawn as a continuous line in figure 12. The observed values of δ_n, i.e. the values $\delta_n(\text{obs.}) = \Delta_{rn}(\text{obs.})/\Delta_\infty(\text{obs.})$ are also indicated in the diagram and connected by the dotted lines.

The agreement between experiment and theory appears to be satisfactory. It appears therefore that the correlations exhibited by the results of the present investigation are entirely due to the macroscopic electrical characteristics of the network used, as supposed in the theory. This was, of course, to be expected because of the limited band-widths which had to be employed. It would be valuable indeed to extend these investigations to much greater band-widths, which would perhaps reveal some differences in the statistical character of the fluctuations from shot and thermal sources due to differences in the correlation between the primary events responsible for these effects. Definite conclusions on this problem will therefore have to be postponed until further and more extensive experimental material on this subject becomes available.

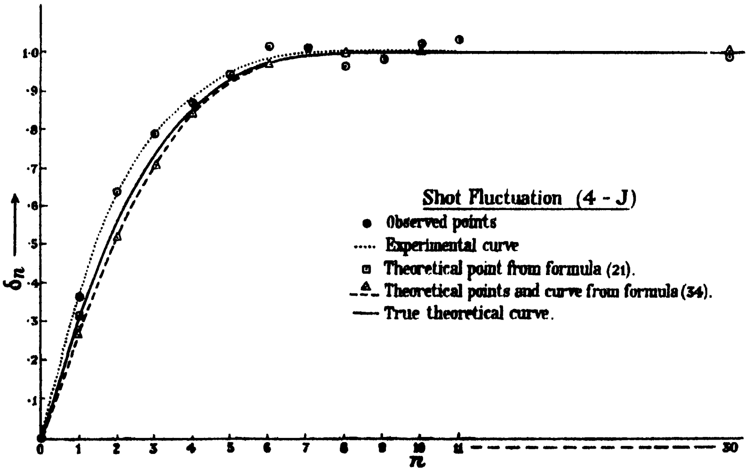


Figure 12. Statistical distribution of δ_n for shot fluctuation, compared with theory.

REFERENCES

BARNES, R. BOWLING and SILVERMAN, S. 1934. *Rev. Mod. Phys.*, **6**, 162.

*CHANDRASEKHAR, S., 1943. *Rev. Mod. Phys.*, **15**, 1.

*DE HAAS-LORENTZ, Mrs. G. L., 1913. *Die Brownsche Bewegung und einige verwandte Erscheinungen* (Braunschweig).

EINSTEIN, A., 1906. *Ann. Phys., Lpz.*, **19**, 380.

*EINSTEIN, A., 1922. "Untersuchungen über die Theorie der Brownschen Bewegung" (Leipzig: Ostwald's *Klassiker der exakten Naturwissenschaften*, Nr. 199); 1926. *Investigations on the theory of the Brownian Movement* (London: Methuen).

FÜRTH, R., 1917. *Ann. Phys., Lpz.*, **53**, 177.

*FÜRTH, R., 1920. *Schwankungserscheinungen in der Physik* (Braunschweig: Sammlung Vieweg).

FÜRTH, R. and MACDONALD, D. K. C., 1946. *Nature, Lond.*, **157**, 807.

JOHNSON, J. B., 1925. *Phys. Rev.*, **26**, 71.

KAPPLER, E., 1931. *Ann. Phys., Lpz.*, **11**, 233; 1932. *Ibid.*, **15**, 545.

MOULLIN, E. B., 1938. *Spontaneous Fluctuations of Voltage* (Oxford University Press).

ORNSTEIN, L. S., 1927. *Z. Phys.*, **41**, 848.

ORNSTEIN, BURGER and TAYLOR, 1927. *Proc. Roy. Soc.*, A, **115**, 391.

* These references deal generally with the subject.

ORNSTEIN, L. S., and MILATZ, J. M. W., 1938. *Physica*, **5**, 971.
RICE, S. O., 1944. *Bell Syst. Techn. J.*, **23**, 282 ; 1945. *Ibid.*, **25**, 46.
*v. SMOLUCHOWSKI, M., 1923. "Abhandlungen über die Brownsche Bewegung und verwandte Erscheinungen" (Leipzig : Ostwald's *Klassiker der exakten Naturwissenschaften*, Nr. 207).
*UHLENBECK, G. E. and ORNSTEIN, L. S., 1930. *Phys. Rev.*, **36**, 823.
*WANG, M. E. and UHLENBECK, G. E., 1945. *Rev. Mod. Phys.*, **17**, 323.
WHITTAKER, E. T. and WATSON, G. N., 1935. *Modern Analysis* (Cambridge : The University Press), p. 384.

* These references deal generally with the subject

DISCUSSION

on the foregoing papers by G. G. MACFARLANE (p. 366) and R. FÜRTH and D. K. C. MACDONALD (pp. 375 and 388).

Mr. D. A. BELL. As the paper by MacDonald and Fürth is concerned with a special case where " shot noise " and " Johnson noise " formulae are easily made to give the same answer, I propose to review the general relationship between these phenomena. They are in fact identical to the extent that both arise from the transport of electric current over a given path by a number of discrete units of charge travelling over limited paths, according to the equivalence

$$l \times i \equiv \sum_{q} \sum_{\Delta x} q \frac{\Delta x}{\Delta t} \qquad \ldots\ldots(1)$$

(l=length of path over which current i flows, q=magnitude of an individual charge, Δx=length parallel to l of an individual path, Δt=time which q takes to travel Δx, and the double summation is to cover all charges present in the system and all paths traversed by each charge in unit time). In most cases, the current can be divided into a mean component and a superimposed fluctuation, corresponding to a division of $\Sigma \Delta x/\Delta t$ into a drift velocity and a thermal agitation velocity. Applying this to a metallic conductor, the quantities e/m of the electron, length Δx of the free path and number N of electrons present can be eliminated in terms of the resistance R, which is a measure of the drift component of the double summation on the R.H.S. of (1) ; and assuming an equipartition value $3k\theta/2$ for the total energy of each electron, we find that the mean square fluctuation of current is $\overline{i^2}=(4k\theta/R)df$ (Bell, 1938), in accordance with Nyquists's formula. The proof can also be obtained with any other distribution of electronic energy (Bakker and Heller, 1939). Turning now to the temperature-limited diode, $q=e$, Δx=cathode to anode distance, Δt is vanishingly small, N is known from the mean current (instead of being measured by a resistance), and we obtain Schottky's formula $\overline{i^2}=2ie\,df$. In the space-charge-limited diode, Δx is still constant ; but Δt, the transit time, is a function both of the anode potential V_a and of the initial thermal energy of the electrons, which has a part $k\theta/e$ electron volts corresponding to the component of velocity directed towards the anode. The relation between $\overline{i^2}$ and i in a space-charge-limited valve is then a function of the ratio $eV_a/k\theta$ (Williams, 1941). In the space-charge-limited diode, Δx is the electrode spacing, N is known from the mean current, and Δt can be calculated for plane and for cylindrical diodes (Bell, 1942) : it is the transit time making due allowance for the initial velocities. One can then calculate the fluctuation current corresponding to any given values of mean current and $eV_a k/\theta$, on the assumption that the thermal components of velocity are responsible for the fluctuation, the mean current being maintained constant by the controlling effect of the space-charge and potential barrier. Pierce has pointed out (Pierce: in Bell, 1943) that by regarding as effective only the difference between the individual thermal velocities and the mean forward component of all thermal velocities, my method of calculation is equivalent to the diode having Johnson noise with a temperature equal to 0·644 times cathode temperature, which is in agreement with other theoretical calculations (Rack, 1938). On the other hand, using the full magnitude of thermal velocities in the formula for cylindrical diodes, I get results in reasonable agreement with published measurements of noise in diodes. The retarding-field condition, which is investigated by the authors, is unique in that agreement of the shot noise expression with the Johnson noise expression requires the electron stream to have a temperature exactly half the cathode temperature. This is a case which I have not worked out from first principles ; but it may be said that the only practical way of observing the temperature of the electron stream is by measuring the noise, since it is only through the noise that we can get that interchange of energy between the electron stream and an external body which is necessary for thermometry. With this proviso, there can be no disputing that the noise generated in the current stream of a valve having a finite internal resistance may properly be expressed as a thermal noise conforming to Nyquist's theorem.

As regards Dr. Macfarlane's paper, it was stated that the apparent relaxation time, which appears in the relation between noise amplitude and frequency, corresponded to a periodicity of the order of one cycle per hour. Is it suggested that, in the thermionic flicker effect for example, the diffusion of ions proceeds at this sort of speed ? I am not satisfied that the noise in poor conductors, such as carbon, is likely to be of the same origin as in barrier-layer semi-conductors having rectifying and photoelectric properties which are known to depend on phenomena on an atomic scale. Could not the noise in carbon resistors arise from macroscopic contact variations at the boundaries between granules, not involving the formation and loss of ions ? Again, the linear increase of noise with period of the measuring apparatus seems to me to indicate merely some degree of positive correlation between current density and conductivity : any fluctuation then tends to be self-maintaining, so that long pulses are more probable than short pulses. The mechanism of such a correlation may differ in different substances.

Dr. C. S. **Bull.** The strong resemblance between the equations for the noise of resistances and that of some valves tempts one to speculate on the connection between the two phenomena.

Nyquist's equation is derived, and is valid only, for a resistance of any kind in thermal equilibrium with its surroundings. Such equilibrium permits the random variation of the energy of the resistance by sharing its energy with the much larger amount contained in the constant temperature surroundings. The energy of the resistive circuit itself is therefore not conserved.

In a valve, however, except in one special case, there never exists thermal equilibrium with the surroundings. In fact, it is easy to follow the energy exchanges arising from the passage of electrons across the circuit capacity, the loss of energy from the cathode due to the emission of electrons, and the power expended by the noise voltage in complete detail by applying the principle of conservation of energy to the valve and its circuit. The result of such an investigation produces the well known equations for the full shot noise and for the noise of a valve in the retarding region of the characteristic. The principles applied are so different from those used in deriving Nyquists' equation that the resemblance of some of the resulting equations does nothing to indicate any possible connection between the two processes.

Turning to the special case discussed by D. O. North in his papers on noise, about 1940, we find that he contemplated a valve with two isolated electrodes in a constant temperature enclosure, one electrode (the cathode) having a lower work function than the other (the anode). It is easy to show that an impedance exists between these electrodes, and that the full shot noise of the equal and opposite currents from cathode to anode and anode to cathode generates the noise voltage given by Nyquist's equation for this valve. This result, of course, could have been expected, since Nyquist's theory puts no restriction on the nature of the resistance. At this point, North removes the anode-cathode current, thereby converting the system into an ordinary valve in the retarding region, and obtaining the well known equation for the noise of such a valve. Physically, such a change in the anode current can only be brought about by the destruction of the thermal equilibrium so necessary to the application of Nyquists' equation to the valve. North's special valve therefore fails also to demonstrate any possible connection between the noise fluctuations of valves and circuits.

Mr. D. A. **Wright.** The physical processes envisaged by Dr. Macfarlane as accompanying current flow through the semi-conductor lead to a law relating noise, current density and frequency which has considerable experimental support. I should like to point out, however, that there are some difficulties associated with this picture. It is essential to the argument that clusters of atoms migrate to the anode side of the semi-conductor and there spread over the surface. Before their arrival, electrons were not readily emitted from that part of the sufrace, but following their arrival, and until their loss by diffusion away as ions, the electron emission is enhanced. Now there are several arguments, which we cannot develop here, against the view commonly held that emission from an oxide coating is dependent on barium atoms adsorbed on the outer coating surface. However, in the case of a semi-conductor with direct contacts, the difficulties become quite clear. The theory requires that the flow through the semi-conductor is controlled by the transfer of electrons from semi-conductor to metal at the positive electrode. When adsorbed atoms are present on the surface of the semi-conductor, this transfer is increased. Thus the theory supposes there is no limitation of flow at the negative electrode where electrons pass from metal to semi-conductor. Thus when the system has rectifying properties, the direction of difficult flow for electrons must correspond with electron flow from semi-conductor to metal. This is contrary to the known direction of rectification in the case of an excess semi-conductor such as zinc oxide. This objection would therefore apply equally to barium oxide.

To conform with the experimental results and theories of contacts between metals and semi-conductors it is necessary to suppose that the current flow is determined at the boundary where electrons pass from metal to semi-conductor, and we might expect the noise to be determined similarly by conditions at this boundary. While details of the arrival and loss of ions and atoms at this boundary are probably very important, it is not easy to see how cluster movement could occur, and it is therefore not easy to apply Dr. Macfarlane's theory, at least if we confine our attention to the movements of the metal ions and atoms. It is just possible that oxygen movement might occur in clusters. In the absence of clusters, the theory would predict again a law of the Schottky type.

Mr. Z. JELONEK. When a theoretical formula is checked experimentally, it is important to consider what extent of agreement can be expected. The correlation functions of the noise amplitude derived theoretically and from the experimental record show good agreement. In this connection I should like to enquire, firstly, how and with what accuracy the Fourier transformation necessary to derive the correlation function from the frequency response of the amplifier network was obtained. Secondly, what was the expected accuracy of the derivation of the correlation function from the noise record. It is well known that in order to obtain good accuracy a very great number of readings must be taken. For example, if a random variable has Gaussian probability distribution, and if an accuracy about $\pm 0{\cdot}01$ in the product moment correlation coefficient is required, it may be necessary to take as many as 10,000 readings. (See, for example, *The Advanced Theory of Statistics*, by M. G. Kendall, vol. i, p. 211.)

In view of such consideration I am interested to know whether the procedure used was sufficiently detailed to make the good agreement shown really significant.

Mr. R. DEHN. I should like to ask Dr. Macfarlane if any reason can be given for the *decrease* of the described type of noise with *increasing* temperature.

Mr. B. N. WATTS. I should like to ask Dr. Macfarlane if there is a correlation between the value of x in his formula z^{x+1}/f^x and the method of deposition, and if it is affected by the presence of the photo-conductivity effect in the film.

Dr. M. PIRENNE. I should like to ask Dr. Fürth and Dr. MacDonald whether their experiments give information about statistical fluctuations in the number of light quanta acting upon the photocell. Is the outcome of the experiments the same as if the number of quanta acting upon the photocell, and the number of photo-electrons liberated, were perfectly constant in time ?

When variations in the nominal light intensity are produced by the experimenter, these artificial variations of course can have an effect on the electric current of the apparatus. Can the natural fluctuations of intensity—due to the quantum nature of light—alter the characteristics of the electric current, or not ? If the case were comparable to that of a Geiger-Müller counter for X-ray photons, these fluctuations should affect the current.

Dr. E. G. JAMES. Dr. Macfarlane has stated that the migration of clusters of atoms to the surface of the cathode coating accounts for the low-frequency noise in thermionic valves, and that the same mechanism would explain the excess noise found in the other semi-conductors. The " flicker " effect associated with oxide cathodes is only noticed in measurements made at audio- or very low radio-frequency, while in certain semi-conductors, e.g. a tungsten-silicon rectifier, excess noise is found to exist at frequencies up to several megacycles per sec. Is it possible to explain this difference on the above theory ?

Another possible mechanism which would give rise to excess noise at low frequency is the movement of ions across the potential barrier between the metal and semi-conductor. These ions would either increase or depress the potential at the barrier, and would cause a fluctuation in the current, the frequency spectrum of which would depend on the mobility of the ions.

As Mr. Wright has mentioned, there is evidence of a potential barrier between the cathode coating and the cathode core in a thermionic valve, and the fluctuation of this barrier may be the source of the low-frequency noise in this case also.

Dr. G. G. MACFARLANE. In reply to Mr. Bell, the figure of one hour for the relaxation time refers to a semi-conductor such as lead sulphide at room temperature. In a thermionic emitter at, say, 950° K. it is of the order of 1/400 sec. (Sproull, 1945). This corresponds to the average time taken to ionize a cluster of impurity atoms which have diffused on to the surface of a contact. It does not by itself tell us anything about the rate of diffusion of ions, since we do not know how far an ion must be from the surface before its dipolar moment ceases to influence the work function.

Regarding Mr. Wright's objection to the mechanism proposed, it is important to differentiate between the total flow of electrons and the noise fluctuations. In the case of an excess semi-conductor with direct contacts to metal it is true that the mean current is largely controlled by the contact where electrons flow from metal to semi-conductor, but fluctuations can still occur due to adatoms at the other contact, provided these alter the work function at the contact. The magnitude of the fluctuations will be smaller than if the clusters of adatoms occurred at the high-resistance contact, but the frequency spectrum will be the same.

In reply to Mr. Dehn, I would like to correct any impression I may have given that the noise increases with temperature without any further restriction. The statement is only true in the theory provided the mean current is kept constant. This behaviour is expected for the following reasons :—

(i) In the formula for $\overline{\Delta j_2}$ the factors most sensitive to temperature are q^{2m} and $1/\omega^{2m}$.

(ii) The index m is a function of a, which, being Langmuir's constant, is proportional to $1/T$. Therefore since m increases as a is reduced, it also increases as T is increased.

(iii) The decay constant q depends on T through a factor of the form $\exp(-T_0/T)$, as shown on page 373. Therefore we can write

$$\begin{aligned}\overline{\Delta j_2}/j_1 &\propto q^{2m}/\omega^{2m}\\ &\propto \exp\{-2m(T_0/T+\log\omega)\}.\end{aligned}$$

In view of the increase of m with T, the index $2m(T_0/T+\log\omega)$ is found to be a function that increases with T, so that the normalized power density of the noise decreases with rise of temperature.

In reply to Mr. Watts, I do not know of any definite correlation between the index x and either the method of preparation or the photo-conductivity.

I am not sure whether the excess noise in tungsten-silicon rectifiers mentioned by Mr. James shows the type of frequency spectrum associated with flicker noise. Noise in excess of Johnson noise does occur at these frequencies in most semi-conductors, due to fluctuations in the number of electrons in the conduction band, but its spectral power density does not fall off rapidly with increasing frequency.

Dr. R. FÜRTH. Several attempts have been made in the past to prove the existence of fluctuations in radiation by measuring fluctuations in the emission current of photoelectric cells, as suggested by Dr. Pirenne. But the objection to such experiments is that a photocell must exhibit fluctuations even in the complete absence of radiation fluctuations, owing to the fact that the relation between the incoming photons and the emitted electrons is of a statistical nature. More recently one of my collaborators (A. Kolin, *Ann. Phys., Lpz.*, **21**, 813 (1933)) used an ultra-violet counter for a similar experiment (i.e. a Geiger-Müller counter with an ultra-violet sensitive cylinder) which permits one to count ultra-violet photons separately. However, it appeared from a statistical analysis of the time-interval distributions of the counts that here too the effect of the fluctuations in the number of incoming photons is completely masked by the efficiency fluctuations of the counter. We may therefore conclude that the effect of statistical fluctuations of the illuminating light on the experiments of Macdonald can be disregarded.

Dr. D. K. C. MacDonald. In reply to Mr. Jelonek, the Fourier transformation involved in the Wiener-Khintchine theorem was obtained on the assumption that the overall frequency-response of the amplifier network was Gaussian. This is a customary assumption for a network of relatively narrow band-width composed of a number of tuned circuits in cascade, which agrees well in practice. The Gaussian parameter was determined by numerical integration of the response curves ; it was felt that this method (rather than, for example, logarithmic plotting of the response curve) would best give effect to the whole response curve.

As regards the second point, it should be emphasized that in our work a *difference* correlation-coefficient was employed throughout. Consequently, for the larger values of correlation-interval, where the examination of the agreement between theory and experiment becomes particularly interesting, the absolute magnitude of the correlation coefficient is relatively large (tending to a constant value for infinite interval) ; therefore the proportionate accuracy obtainable is quite adequate with a reasonable number of observations. We may further mention that 200–300 primary observations were used on each record, and in all some eight records were examined in statistical detail ; we were entirely satisfied that the good agreement obtained overall and in individual records was significant.

In confirming Dr. Fürth's remarks relative to Dr. Pirenne's question, I should also like to draw attention to a similar problem relating to the voltage fluctuations of a piezo-electric crystal discussed recently elsewhere (Lawson and Long, 1946; Brown and MacDonald, 1946). In that case the relationship of the electrical fluctuations to the concomitant spontaneous mechanical vibrations was under review ; it was there emphasized that in thermal equilibrium the magnitude of the electrical fluctuations generated is completely determined by the electrical characteristics of the system (Nyquist's theorem) independent of the particular electro-mechanical interaction involved.

As regards the points raised by Mr. D. A. Bell and Dr. C. S. Ball, it is, in my opinion, evident that in general Nyquist's theorem cannot be applied directly to a thermionic valve as a whole to determine the noise generated at the external terminals of the valve. It does, however, appear to me entirely reasonable to ascribe to the cathode-potential-barrier region *generally* a *short-circuit* noise current given by

$$\overline{(i-I)^2_b}=\tfrac{1}{2}(4g_b kT\Delta f), \qquad \ldots\ldots(1)$$

where g_b is the barrier-cathode differential conductance. On applying Helmholtz's (Thévenin's) constant-current theorem, the short-circuit noise current obtaining at the external (cathode-anode) terminals will be less than (1) by a formally simple impedance transformation factor involving g_b and the barrier-anode impedance. Noting that in the parallel-plane structure $g_b=\frac{eI}{kT}$, it follows that the " space-charge reduction factor ", Γ^2, is fundamentally no more than this impedance transformation. Schottky (1936) formulated the impedance-transformation concept in a somewhat lengthy discussion, but was in error in evaluating it. It is clear (MacDonald, 1946) that when the barrier : anode impedance is evaluated on the basis of the Maxwell-Boltzmann law, this view-point leads to precisely the same result as that obtained by North (1940), Rack (1938) and Schottky in his later work (1937). The actual numerical computation (as undertaken by these workers) is naturally rather complex, but the same is true of the evaluation of any other theoretical valve parameter when account is taken of the velocity emission law, and it should not be inferred that the noise problem itself presents any greater difficulty in principle. In this sense, and to this extent then, I believe that " shot " and " thermal " noise may be unified.

REFERENCES

Bakker, C. J. and Heller, G., 1939. *Physica*, **6**, 262.
Bell, D. A., 1938. *J. Instn. Elect. Engrs.*, **82**, 522.
Bell, D. A., 1942. *J. Instn. Elect. Engrs.*, **89**, 207.
Brown, J. B. and MacDonald, D. K. C., 1946. *Phys. Rev.*, **70**, 976.
Lawson, A. W. and Long, E. A., 1946. *Phys. Rev.*, **70**, 220, 977.
MacDonald, D. K. C., 1941. Ph.D. Thesis, Cap. 2, pp. 31–2 (University of Edinburgh).
North, D. O. *et al.*, 1940. *R.A.C. Review*, **4**.
Pierce, J. R., 1943. (Discussion on Bell, 1942.) *J. Instn. Elect. Engrs.*, **90**, 148.
Rack, A. J., 1938. *Bell Syst. Tech. J.*, **17**, 592.
Schottky, W. *et al.*, 1937. *Wiss. Veröff. Siemens-Werk.*, **16**, 1.
Schottky, W., 1938. *Die Tel. Röhre*, **8**, 175.
Williams, F. C., 1941. *J. Instn. Elect. Engrs.*, **88**, 219.

THE MASS OF THE NEUTRINO

By F. C. FRANK,
H. H. Wills Physical Laboratory, Bristol

MS. received 6 *February* 1946; *in revised form* 16 *December* 1946

ABSTRACT. From a collation of the energies involved in nuclear reactions and radioactive decays, which together represent closed cycles, it is shown that the rest-mass of the neutrino, m_ν, is beyond any reasonable doubt less than that of the electron, m_e, and probably less than $m_e/10$. It cannot be shown with certainty to be measurably distinct from zero.

§ 1. INTRODUCTION

In Mattauch and Flügge's *Kernphysikalische Tabellen* (1942) there is only one statement concerning the rest-mass of the neutrino, viz. that on p. 76 that it is certainly not larger, and is probably smaller, than that of an electron. This, however, is an almost tendentiously conservative statement: it is clearly a question of considerable significance whether the mass of the neutrino, m_ν, is of similar magnitude to that of the electron, m_e, or no. Such experimental evidence as we possess, which is presented elsewhere in Mattauch and Flügge's Tables, indicates fairly unambiguously that it is not. This has been the usual view adopted in theoretical considerations of the matter. F. Perrin (1933) surmised on qualitative grounds that the rest-mass of the neutrino ought to be zero to account for the average value of the β-energy. Bethe and Bacher (1940) deduced from the form of the β-spectrum, in connection with Fermi's theory (1934), that the neutrino mass is not more than one-fifth that of the electron. Bethe and Bacher also reached a similar conclusion from another argument, based on the number of existing pairs of adjacent odd isobars (i.e. nuclei of the same odd mass number, differing in atomic number by 1). It is observed that the energies of odd isobars, plotted against atomic number, lie close to smooth parabolae, the form of which depends on mass number in a known manner. Thus it is possible to make a statistical estimate of the chance that for some odd mass number in the range of the natural nuclei there exists a pair of odd isobars, being the two most stable nuclei for their mass number, differing in energy by less than any stated amount.

It also appears certain that any nucleus will sooner or later undergo β-decay (positive or negative) or will capture one of its planetary electrons, if energy can be released in the process. It is postulated that each such process involves the emission of a neutrino. Then the two adjacent isobars will both be stable if their energies differ by less than the energy equivalent of m_ν. If $m_\nu = 0$, there will be no stability range at all between the reverse processes of electron emission and electron capture, and therefore no stable pairs of adjacent isobars. In fact, there are three pairs of adjacent odd isobars not known to undergo change, viz. Cd—In, mass number 113; In—Sn, 115; and Sb—Te, 123. A fourth pair, Re—Os, 187, listed by Mattauch and Flügge, has since been eliminated (Lougher and Rowlands, 1944). If it is assumed that these three pairs are genuinely stable, for the reason mentioned, we have the statistical estimate $m_\nu c^2 \approx (0{\cdot}03 \pm 0{\cdot}02)$ Mev. If m_ν were as large as m_e, we should expect about 50 such pairs. Thus this argument definitely indicates that the mass of the neutrino is much less than that of the electron. On the other hand a similar calculation shows that there is a statistical likelihood of about two pairs differing in energy by less than the binding energy of a K-electron ($13{\cdot}6\ Z^2$ electron volts). For such nuclei, only electrons from remoter shells are available for capture, and the capture-decay is bound to be slow and difficult to detect. Flügge, in the tables mentioned, summarizes reasons, based on isotopic abundances, for supposing that such undetected changes may occur in these cases. Hence this argument does not exclude the value $m_\nu = 0$.

All of the above arguments are somewhat oblique, and this paper is primarily concerned with more nearly direct estimates of m_ν, from the energies of nuclear processes, using the data available in Mattauch and Flügge's Tables.

§ 2. CYCLE ENERGIES

The mass of the neutrino can be found most directly from the energy balance of simple cycles of nuclear reaction followed by β-radioactive decay, i.e., by the generation of a positive or negative electron, and simultaneously, according to the hypothesis, of a neutrino. There may also, sometimes, be γ-emission to complete the cycle. The simplest types of cycle are accordingly:

$$\text{(I)}\quad {}^{y}_{x}At\,(n,p)\,{}^{\ y}_{x-1}At'\,(\text{-},\, e^-\nu\gamma)\,{}^{y}_{x}At.$$

$$\text{(II)}\quad {}^{y}_{x}At\,(p,n)\,{}^{\ y}_{x+1}At'\,(\text{-},\, e^+\nu\gamma)\,{}^{y}_{x}At.$$

The notation employed here corresponds to that of Mattauch and Flügge. ${}^{y}_{x}At$ signifies the atomic species At, of atomic number $Z = x$ and mass number $A = y$. Brackets signify a nuclear reaction, wherein what appears before the comma enters the nucleus, and what appears after the comma emerges from it. A hyphen before the comma signifies a spontaneous reaction—i.e. a radioactive decay.

If $Q_{(n,\,p)}$ or $Q_{(p,\,n)}$ is the energy yield of the corresponding nuclear reaction stage of the cycle, Q_β is the limiting energy of the β rays and Q_γ the energy of the γ rays emitted, if any, the mass of the neutrino is to be calculated as

$$m_\nu = m_n - m_p - m_e - \frac{1}{c^2}\,(Q_{(n,\,p)} + Q_{\beta^-} + Q_\gamma), \qquad \ldots\ldots(1)$$

$$m_\nu = m_p - m_n - m_e - \frac{1}{c}\,(Q_{(p,\,n)} + Q_{\beta^+} + Q_\gamma). \qquad \ldots\ldots(2)$$

There are additional cases in which, although the nuclear reaction (n, p) or (p, n) is not realized directly, its energy yield can be determined from the two reactions

$$ {}^{y-1}_{x}At''(d, p)\,{}^{y}_{x}At \quad \text{and} \quad {}^{y-1}_{x}At''(d, n)\,{}^{y}_{x+1}At', $$

i.e. reactions in which bombardment with deuterons causes emission of protons and neutrons respectively. Cycles thus realized indirectly can be called types (I *a*) and (II *a*).

Cycles which are completed by electron capture instead of radioactive emission (type III):

$$ {}^{y}_{x}At\,(p, n)\,{}^{y}_{x+1}At'\,(e_{K,}\,\nu\gamma)\,{}^{y}_{x}At, $$

only allow us to find an upper limit to m_ν, since there is no means of measuring the energy carried away by the neutrino,

$$ m_\nu \leqslant m_p - m_n + m_e - \frac{1}{c^2}(Q_{(p,\,n)} + Q_\gamma + Q_x). \qquad \ldots\ldots(3) $$

where Q_x is the energy appearing as X rays when external electrons fill the place of the K-electrons absorbed.

The difference $(m_n - m_p)$ could in principle be eliminated by combining equations (1), and (2), but this is unnecessary and wasteful of measurements, as it is known to a higher accuracy than cycle energies, from mass spectroscopy combined with measurement of the energy involved in the reaction $D(\gamma, n)H$. From Mattauch's best selected data it is

$$ m_n - m_p = 1{\cdot}361 \pm 0{\cdot}025 \text{ TMU}. $$

(Mass is quoted throughout this note in TMU—thousandths of a mass unit, called TME by Mattauch and Flügge; 1 TMU is 1/16,000 of the mass of a neutral atom of ${}^{16}_{8}O$, and its equivalent energy according to $E = mc^2$ is 0·931 Mev.) The mass of the electron, m_e, is 0·547 TMU, and should be free from error to this number of significant figures. In terms of energy,

$$ \begin{aligned} m_e c^2 &= 0{\cdot}509 \text{ Mev.}, \\ (m_n - m_p)\,c^2 &= 1{\cdot}267 \pm 0{\cdot}024 \text{ Mev.}, \\ (m_n - m_p - m_e)\,c^2 &= 0{\cdot}758 \pm 0{\cdot}024 \text{ Mev.}, \\ (m_p - m_n - m_e)\,c^2 &= -1{\cdot}776 \pm 0{\cdot}024 \text{ Mev.} \end{aligned} $$

Table 1 summarizes data for all cycles of the specified types for which there is information in Mattauch and Flügge, Tables IV and VI. Table IV contains, *inter alia*, the limiting energies of β-spectra, which are here entered in column 3. In one case (the β-decay of ${}^{20}_{9}F$) the energy of a γ ray has been added in as well. Table VI collects together measurements of the energy yields of nuclear reactions, which are here entered in column 2. These include some values derived by multiplying the threshold energy for the reaction (p, n) by $A/(A+1)$. In both columns, single weighted mean values are entered in the few cases in which the tables list two or more measurements of comparable precision: in such cases no weight is given to measurements with no declared probable error. The remaining two columns of table 1 are calculated from columns 2 and 3.

Table 1

Nuclei involved in cycle	Q reaction (Mev.)	Q decay (Mev.)	Q cycle (Mev.)	$m_\nu c^2$ (Mev.)
Type I : $(n, p)-\beta^-$				
${}^{14}_{7}\mathrm{N}$ ${}^{14}_{6}\mathrm{C}$	0·55±0·03	0·12±0·02	0·67±0·037	0·088±0·05
Type I *a* : the same from $(d, p)-(d, n)$				
${}^{3}_{2}\mathrm{He}$ ${}^{3}_{1}\mathrm{H}$	0·67±0·036	0·01±0·002	0·68±0·036	0·078±0·043
(from ${}^{2}_{1}\mathrm{D}$) (3·98±0·02−3·31±0·03)				
${}^{10}_{5}\mathrm{B}$ ${}^{10}_{4}\mathrm{Be}$	0·35	0·55	0·90	−0·142
(from ${}^{9}_{4}\mathrm{Be}$) (4·55−4·20)				
${}^{20}_{10}\mathrm{Ne}$ ${}^{20}_{9}\mathrm{F}$	−6·5	7·2	0·7	0·058
(from ${}^{19}_{9}\mathrm{F}$) (4·3 −10·8±0·2) (5·0±2·2)				
Type II : $(p, n)-\beta^+$				
${}^{10}_{5}\mathrm{B}$ ${}^{10}_{6}\mathrm{C}$	−5·1	3·36±0·1	−1·74	−0·036
${}^{11}_{5}\mathrm{B}$ ${}^{11}_{6}\mathrm{C}$	−2·72±0·01	0·981±0·005	−1·739±0·011	−0·037±0·03
${}^{13}_{6}\mathrm{C}$ ${}^{13}_{7}\mathrm{N}$	−2·97±0·03	1·21±0·01	−1·76±0·032	−0·018±0·04
${}^{18}_{8}\mathrm{O}$ ${}^{18}_{9}\mathrm{F}$	−2·42±0·04	0·7	−1·72	−0·056
${}^{19}_{9}\mathrm{F}$ ${}^{19}_{10}\mathrm{Ne}$	−3·97	2·20	−1·77	−0·006
${}^{23}_{11}\mathrm{Na}$ ${}^{23}_{12}\mathrm{Mg}$	−4·58±0·3	2·82	−1·76	−0·016
${}^{27}_{13}\mathrm{Al}$ ${}^{27}_{14}\mathrm{Si}$	−5·8±0·1	3·64±0·1	−2·16±0·14	0·384±0·15
${}^{61}_{28}\mathrm{Ni}$ ${}^{61}_{29}\mathrm{Cu}$	−3·0	0·94	−2·06	0·284
${}^{64}_{28}\mathrm{Ni}$ ${}^{64}_{29}\mathrm{Cu}$	−2·5	0·655±0·003	−1·845	0·069
${}^{63}_{29}\mathrm{Cu}$ ${}^{63}_{30}\mathrm{Zn}$	−4·0	2·32+0·005	−1·68	−0·096
${}^{68}_{30}\mathrm{Zn}$ ${}^{68}_{31}\mathrm{Ga}$	−3·6	1·85	−1·75	−0·026
Type II *a* : the same from $(d, n)-(d, p)$				
${}^{11}_{5}\mathrm{E}$ ${}^{11}_{6}\mathrm{C}$	−3·06	0·981±0·005	−2·079	0·303
(from ${}^{10}_{5}\mathrm{B}$) (6·08−9·14±0·06)				
${}^{13}_{6}\mathrm{C}$ ${}^{13}_{7}\mathrm{N}$	−2·96±0·06	1·212±0·004	−1·748±0·06	−0·028+0·065
(from ${}^{12}_{6}\mathrm{C}$) (−0·25±0·03−2·71±0·05)				
${}^{17}_{8}\mathrm{O}$ ${}^{17}_{9}\mathrm{F}$	−3·65	2·1	−1·55	−0·226
(from ${}^{16}_{8}\mathrm{O}$) (−1·7−1·95±0·06)				
Type III : $(p, n)-K$ (giving upper limit only)				
${}^{7}_{3}\mathrm{Li}$ ${}^{7}_{4}\mathrm{Be}$	−1·62±0·02	0·425±0·025	−1·195±0·032	⩽0·437±0·04

Consistency data ("probable errors") have been included in the table where they are listed, and carried through to the final column, but they have not been utilized for the purpose of weighting the results. Various authors adopt various conventions about the inclusion in the "probable error" of uncertainty in basic calibrations: in general they are a good index of the scatter in measurements, but a poor index of the possible systematic error. Accordingly the 18 cases of types I, I *a*, II and II *a*, have been treated as a homogeneous set of measurements, with a Gaussian expectation of errors. The probable error for each cycle is thus found to be 0·105 Mev., and the mean result with probable error (incorporating the probable error of $m_n - m_p$)

$$\begin{aligned} m_\nu c^2 &= 0{\cdot}0321 \pm 0{\cdot}0357 \text{ Mev.,} \\ m_\nu &= 0{\cdot}0345 \pm 0{\cdot}0383 \text{ TMU} \\ &= (0{\cdot}0631 \pm 0{\cdot}0701)\, m_e. \end{aligned}$$

This statement of "probable error" is subject to the same proviso as that made above—it assumes that errors are unbiased. Two probable sources of bias—errors in locating the limit of a β-ray continuum, and failure to detect a soft γ ray—are both liable to lead to an over-estimate, rather than an under-estimate, of the neutrino mass: a third, over-estimate of the threshold energy for (p, n), would have the opposite effect. This latter error must be larger than all estimated probable errors if the mass of the neutrino is of similar magnitude to that of the electron. In 19 different cycles there is not one which indicates that m_ν is as large as m_e, and m_e is outside the estimated probable error of m_ν by a large factor. Flügge's statement in the tables by Mattauch and Flügge, p. 76, must therefore be rejected as unreasonably conservative. Considering the other evidence as well, the rest-mass of the neutrino is almost certainly less than one-tenth that of the electron, and, so far as present measurements can show, not appreciably different from zero.

REFERENCES

Bethe and Bacher, 1940. *Rev. Mod. Phys.*, **8**, 189.
Fermi, 1934. *Z. Phys.*, **88**, 161.
Lougher and Rowlands, 1944. *Nature, Lond.*, **153**, 374.
Mattauch and Flügge, 1942. *Kernphysikalische Tabellen* (Berlin : Springer)
Perrin, F., 1933. *C.R. Acad. Sci., Paris*, **197**, 1625.

WHAT EXPERIMENTS ARE NEEDED IN FUNDAMENTAL PHYSICS ?

By R. E. PEIERLS,
University of Birmingham

Read 20 *December* 1946; *MS. received* 10 *January* 1947

ABSTRACT. A brief outline is given of the chief outstanding difficulties of fundamental theory concerning the nature, origin and properties of nuclear forces, the properties of mesons, their inter-relation with other particles and their connection with nuclear forces, the lifetime of the meson and its relation to ordinary beta-processes, with the classification of atomic energy levels. A discussion is given of some experiments which might become possible with modern technique and which are likely to help provide the answer.

§ 1. GENERAL

In discussing the value of possible experiments in fundamental physics from the point of view of present theory, it is not my intention to suggest that it is always possible to predict the kind of experiment that will lead to an important advance in our knowledge. Indeed, from past experience, we must certainly be prepared for further unsuspected discoveries which can only come from a broad study of fundamental phenomena over a wide field. As a typical example of this kind of discovery I need only quote the discovery of x rays; Roentgen certainly set out to study in a general way what happens in a cathode-ray tube, but what precisely he was going to find he did not know in advance. On the other hand we have discoveries like Rutherford's law of the scattering of alpha particles; this was the result of a deliberate attempt to explore the inside of atoms by that particular tool.

It is only experiments of this latter type that are amenable to discussion from a theoretical point of view, but this should not be understood as an attempt to minimize the importance of new and unpredicted observations.

§ 2. NUCLEAR FORCES

Progress in the theory of nuclei is hampered by lack of knowledge of the precise nature of the forces acting between elementary particles. In this respect the situation differs from that of atomic theory, where, as soon as the existence of the electron and the nucleus was discovered, one could take it for granted that the forces between them were electric in nature and were essentially given by Coulomb's law. The difficulty was there in the change of the general laws of mechanics that found its expression in quantum theory.

In the case of the nucleus it is likely that the general laws of quantum mechanics can, to a large extent, be applied to the phenomena inside the nucleus, although there are some doubts on that score, but the nature of the forces is still essentially unknown. The most direct information about the forces must come from the study of two-body problems, in the same way in which, in atomic theory, the hydrogen atom provided the crucial confirmation for Bohr's theory; but while the hydrogen spectrum contains a large amount of simple data, such as the line frequencies, selection rules, Stark and Zeeman effects, etc., the corresponding nuclear two-body problem, i.e. the deuteron, has no excited states and no spectral lines and cannot be influenced to a measurable extent by external fields. We must therefore fall back on the properties of reactions with two bodies, such as the scattering and capture of neutrons by protons and the scattering of protons by protons. It turns out that, given the binding energy of the deuteron and the capture cross-section for thermal neutrons in hydrogen, the behaviour of all collisions of two particles at energies up to a few million volts can be predicted quite well without further assumptions about the details of the law of force. Owing to the fact that these forces act only over very short range, new information can be obtained only from cases in which the de Broglie wave-length is comparable to the range of the forces, and that means at energies of well above 10 Mev. At energies of this order, experiments by Amaldi and his collaborators have tended to show an asymmetry in the scattering which, if confirmed, could be used to decide whether the forces are "ordinary" or "exchange" forces, but the asymmetries in question are very small, and experiments by Powell in Bristol, using a different method, contradict Amaldi's results. Measuring the intensities of neutron beams precisely is always difficult, and more conclusive evidence could be obtained by going to much higher energies, where the expected asymmetries are larger.

In the case of proton-proton scattering, precision measurements are easier and, while the interpretation of the experiments is more difficult because of the presence of the long-range electrical repulsion, in addition to the nuclear forces, the theory of this problem is sufficiently well known to draw definite conclusions as soon as data at energies of the order of 10 Mev. are available. Proton-proton scattering, however, cannot give all the necessary information, since, for example, the distinction between "exchange" and "ordinary" forces does not arise in the case of identical particles.

Experiments at still higher energies, which will be capable of giving information on the forces at extremely short distances, are also of importance since we now know that the law of forces between two nuclear particles contains a directional force rather like the interaction between two electric or magnetic dipoles. The most elementary theory of this directional or "tensor" force leads to a potential varying with the inverse cube of the distance, as in the case of dipoles, and it is known that the singularity at zero distance is then so strong as to make any stationary state of finite binding energy impossible. The law must therefore be more complicated at small radii, and this is just the point where theory meets one of its greatest difficulties.

A different type of problem is associated with those phenomena in which one observes the interaction of protons or neutrons with an external electromagnetic field, such as the capture of neutrons by protons or the photo-dissociation of the deuteron. The particular interest of these problems lies in the fact that, according to the meson theory of the forces, the proton is not just a point charge but spends a fraction of its time as a neutron with a charge spread over a small region in a surrounding field in which it is trying to generate positive mesons. This view is supported by the fact that the magnetic moment of the proton is not just one nuclear magneton (Bohr magneton divided by the proton-electron mass ratio), as Dirac's theory leads one to expect, but considerably larger. This view of the structure of the proton could be confirmed by comparing quantitative results about the emission and absorption of electromagnetic radiation by protons with the theory assuming a point charge.

§ 3. MANY-BODY PROBLEMS

In the problems of heavier nuclei the success of the general picture first proposed by Bohr has shown that a great many features can be qualitatively understood without specific reference to the nature of the forces. It is therefore difficult to use experiments on heavy nuclei for conclusions about the forces. In principle, of course, a quantitative knowledge of the energy levels and other properties of nuclei could be used to test, or even to derive, the quantitative details of the interaction. However, in practice this would be about as difficult as the job of deriving atomic theory from a knowledge of the frequencies and intensities of the lines in the iron spectrum. In a way the problem is even more difficult because, while in the atom the presence of a strong centre of force and the shell structure allow us to treat each electron to a reasonable first approximation as moving in a given field of force, this simplification is of no help in the case of the nucleus. What would, however, be of great assistance is statistical material about the energy levels, symmetry properties, transition probabilities, etc., of the excited levels of very many nuclei. For instance, the fact that a particular nucleus has a large capture cross-section for slow neutrons has very little direct significance, but knowledge of the number of isotopes in a certain part of the periodic system which have cross-sections above a given value can be used to estimate the distribution of excited states of nuclei of that type. In this connection it would be of particular importance to have some studies of nuclei about which as much information as possible is available simultaneously, such as the energy of the level, its angular momentum and parity, the intensity of any gamma-rays and beta-transitions involving this level, information on the internal conversion of the gamma-rays, etc.

In this connection I would like to draw particular attention to the study of excited levels by means of collisions with charged particles (the nuclear analogue of the Franck-Hertz experiments), which needs careful measuring technique but does not require very high energies and which, while unspectacular, would provide most valuable information.

§ 4. BETA-DECAY

Our knowledge of the beta-ray spectrum has improved a great deal and it looks at present as if the simplest form of Fermi's theory accounts for it reasonably well. Even there the form of the theory is not uniquely fixed by the spectrum, but several alternative laws are possible which give the same spectrum but different selection rules and possibly different spectra for "allowed" (i.e. intense) and "forbidden" (i.e. weak) transitions. What is needed here is, therefore, a close study, in suitable cases, of the symmetry type and spin of the initial and final states of the decaying nucleus, and a knowledge of any gamma-rays emitted before or after the beta-ray. The study of the beta-ray spectrum at very low energies seems to be particularly difficult from the point of view of experimental technique, and it would be important to know how far the remaining discrepancies between theory and experiment at the low energy end can be reduced by an improvement in technique.

The important feature of beta-ray theory is the need for assuming a neutrino, and further experiments of the recoil type would put our belief in its existence on a firmer basis and at the same time, by measuring the angular distribution of the neutrinos in relation to the electron, would help to differentiate between possible alternative theories.

I need hardly say that we will never be quite satisfied about the neutrino unless we have succeeded in finding it not only emitted but also absorbed, but for this something entirely new would be required, and this belongs to the type of discovery for which theory can be of little assistance.

§ 5. MESONS

A great advance in our knowledge of fundamental particles can be expected when means become available to produce in the laboratory particles of sufficient energy to generate mesons. Cosmic-ray evidence suggests a meson mass of somewhat over 100 Mev., but it is not clear at present whether all mesons have the same mass or not, and for a full exploration it would be important to have some excess energy in hand. In principle, mesons could be generated as soon as the kinetic energy of the particle exceeds the rest energy of the meson, but, as in the case of pair creation, one would expect that the effective cross-section for this process would be small immediately above the threshold and that reasonable intensities could only be obtained with energies well above the minimum.

As soon as mesons can be made in the laboratory, one will clearly want to study all their properties, including the results of their collisions with nuclei and other particles and processes in which they are absorbed.

At the same time it is important to find out whether, in addition to charged mesons, there exist neutral ones. If the general ideas of the current meson theory of nuclear forces are justified, neutral mesons must exist and must be produced fairly readily in collisions of fast neutrons or protons with each other or with nuclei.

This group of problems is probably the one in which we may expect the most spectacular advances in our fundamental knowledge as soon as machines for generating the necessary energies come into operation.

§ 6. ELECTRONS AND ELECTROMAGNETIC RADIATION

The problems of the properties of electrons and their interaction with the radiation field at very high energies and for collisions involving very close approach are probably the most important ones in connection with the deadlock in present theory. No adequate theory is available for describing such interactions at all and, while certain ways of calculation are available, which in all likelihood give the right answers for such things as the scattering of fast gamma-rays by electrons, the emission of light and the creation of pairs in electronic collisions, etc., they all contain arbitrary procedures. If any of the present theories were carried out with logical consistency, one would in most cases get either zero or infinity as an answer. It is not easy to suggest in detail what experiments are likely to help in finding a solution of these difficulties, except that the more our knowledge of the phenomena involving fast electrons, the greater the chance of bringing order into the present complex situation.

DISCUSSION

Mr. E. S. Shire. May I ask Prof. Peierls at what energies he would expect the current theory of collisions between elementary particles, particularly electrons, to break down ?

Prof. M. L. Oliphant. I should like to ask Prof. Peierls whether any phenomena of the nature of " polarization " phenomena are likely to appear with fundamental particles at close distances of approach ?

Dr. O. Frisch. I agree with Prof. Peierls that much experimental information of a statistical character (level density, distribution of width values etc.) should be collected to increase our knowledge of nuclei. At the same time I want to point out that there are some known experimental facts which indicate the presence of definite structural features even in heavy nuclei. In particular, there are several facts which show that various nuclear properties are very little affected by the precise number of neutrons in the nucleus. The most striking fact of this kind is the strict parallelism of the three well-known natural radioactive series ; they all show alpha decay of increasing violence down to the appropriate lead isotope, when alpha decay suddenly stops, they all show branching at the bismuth isotope, and so on.

Another structural feature shows up if one plots the magnetic moments of nuclei against their angular momenta. Nuclei containing an odd neutron have small magnetic moments, of the order of that of the free neutron, while nuclei containing an odd proton have magnetic moments increasing with the angular momentum. It looks as if the odd nucleon was actually running on an orbit.

Phenomena of that kind raise the hope that a partly structural (and not merely statistical) model of atomic nuclei will eventually emerge.

Prof. P. B. MOON. I would like to call attention to another well known fact that may point to a similarity of structure between nuclei that differ by one or a few neutrons ; namely the occurrence of metastable states in several isotopes of a few elements, such as Te and In.

AUTHOR'S REPLY. With regard to Mr. Shire's question, it is not easy to judge what energies should be regarded as " high " for this purpose. One would guess that the order of magnitude should be that energy at which the de Broglie wave-length becomes comparable with the classical electron radius. This is of the order of 137 electron masses. One must, however, remember that we shall get no new information from processes involving fast particles which, after a suitable Lorentz transformation, appear to depend only on phenomena involving slow particles. It is probably right to say that it is not sufficient to have particles of energies of that order, but also to observe with them phenomena in which a close collision of the particles to distances of the order of 10^{-13} cm. is essentially involved. Such processes will, in general, involve cross-sections of the order of 10^{-26} cm^2 or less, and we can expect interesting information only from conditions in which such comparatively rare events are made observable.

In reply to Prof. Oliphant, while some of the current theories talk of a finite size for elementary particles, I would not like to interpret this as a definite structure, since then relativity would make it impossible to assume the structure as rigid, and one would have to admit internal degrees of freedom of such a structure which, upon quantization, would again lead to new particles. Instead, one likes to regard the structure as connected with a limitation in our fundamental concepts about space which might make it impossible to locate the charge belonging to the particle at a precise mathematical point. In this sense the structure is not liable to change under external forces and the question of polarization does not arise.

All current theories, however, involve the assumption that since the electron can virtually generate light quanta and light quanta can generate pairs of electrons, whereas protons and neutrons can generate mesons, all these virtual processes express themselves in disturbances in the space surrounding each particle. I have referred to one example of this in the case of the charge distribution near a proton. These disturbances may themselves be influenced by the presence of the particles and in that sense there may exist polarization effects. However, just on this point the present theory is on rather uncertain ground and it is one of its unsatisfactory features that a phenomenon apparently so simple as a single electron or light quantum or meson, or even empty space, appears in the theory as a state of affairs of enormous complexity.

The kind of consideration to which Dr. Frisch and Professor Moon have referred revealed general regularities in the properties of nuclei, which, I agree, may very well help in forming a more systematic picture of nuclear properties. Unfortunately no-one has, as yet, succeeded in drawing clear conclusions from them. It is true they are compatible with the idea of alpha particles existing as separate sub-units in the nucleus, but they are equally compatible with the opposite extreme of individual neutrons and protons moving in independent orbits, and neither of these extremes is likely to be near the truth.

COLLISION BROADENING OF THE INVERSION SPECTRUM OF AMMONIA AT CENTIMETRE WAVE-LENGTHS. I.—SELF-BROADENING AT HIGH PRESSURE

By B. BLEANEY and R. P. PENROSE,
Clarendon Laboratory, Oxford

MS. received 3 *January* 1947

ABSTRACT. The absorption spectrum of ammonia between 0·6 and 0·9 cm.$^{-1}$ has been measured at pressures up to 60 cm. Hg. From the previous analysis of the fine structure, and the measurement of the widths of the lines at a pressure of 0·5 mm. Hg, the shape of the absorption curve at high pressures has been computed assuming that the widths vary accurately as the first power of the pressure. At a pressure of 10 cm. Hg, the computed and observed curves agree within the experimental error ; at 60 cm. Hg, the observed attenuation at the lower frequencies is somewhat greater than that computed. Possible reasons for this phenomenon are discussed.

§ 1. INTRODUCTION

The phenomenon of the broadening of spectral lines in an absorbing or emitting gas owing to collisions with other molecules has been studied both in the optical region and in the infra-red. In these regions the contribution to the width of a line due to collision broadening is, at ordinary pressures, of the same order as that due to the Doppler effect. In the region of centimetre wave-lengths, however, the situation is very different. At atmospheric pressure the collision frequency of the molecules of a gas is of the order of 10^{10}, which is comparable with the frequency of the radiation. Spectral lines in this region will therefore be extremely broad at atmospheric pressure. Other contributions to the widths are, on the contrary, very small; the most important is that due to the Doppler effect, which produces a width which is proportional to the frequency of the radiation, and about 10^{-6} times smaller. At pressures down to a few hundredths of a millimetre, therefore, the width of a spectral line at these wave-lengths will be determined solely by collision broadening, and the opportunity arises for studying pressure broadening in some detail. Furthermore, in the centimetre wave-length region, accurate measurements of absorption intensities can be made; thus the shape of a spectral line, and the variation of its width, can be studied over a range of pressures from atmospheric, or greater, down to 0·01 mm. or less. Such pressures have the additional advantage that effects due to multiple encounters should be small, and the complications which these introduce into work at pressures of several atmospheres should therefore be absent.

The only strong resonant absorption at present known to exist at centimetre wave-lengths is that due to the inversion of the ammonia molecule in its ground state. This spectrum, which possesses a considerable fine structure, has been analysed (Bleaney and Penrose, 1946 a, b), using pressures of about one millimetre of mercury. The frequencies of 29 lines were determined, and it was found that they could be represented by a simple formula from which the frequencies of the remaining lines could be calculated. For seventeen of the lines, the intensities and the half-widths were measured, and it was shown that in each case the strength of the line agreed, within the experimental error of about 5%, with the calculated values. Sufficient data are thus available for the use of this spectrum for a study of pressure broadening.

The pressures used in this work (about 1 mm. Hg) were chosen so that the lines were sufficiently narrow for adequate resolution, but still broad enough to permit accurate measurement of their widths. The line-breadth constants * varied between 2.10^{-4} and 5.10^{-4} $\mathrm{cm.}^{-1}$ at a pressure of 0 5 mm. Hg. This pressure lies near the middle of the range of pressures indicated above as suitable for the study of pressure broadening, and it was therefore possible to continue investigation of the spectrum at both higher and lower pressures. In the latter case a complication arises due to the disturbance of thermal equilibrium in the gas; the absorption of energy from the radiation tends to equalize the populations of the upper and lower levels of the transition, and at low pressures the frequency of collisions becomes too small to counteract this process fully and thus maintain the distribution appropriate to the temperature of the gas. The effect has been analysed, and the results, which are in close agreement with the theoretical predictions, will be published shortly in another paper.

The experimental determination of the absorption at higher pressures (up to 60 cm. Hg) had been carried out in 1945 during a preliminary survey of the ammonia spectrum. This survey was mentioned in a précis of the work (Bleaney and Penrose, 1946 a) but was excluded from the more detailed paper (Bleaney and Penrose, 1946 b). The reasons for this were (*a*) the experimental technique was different, (*b*) only a qualitative interpretation of the results was possible before the analysis of the spectrum had been made. In this paper the experimental technique and results are described, and the absorptions at pressures of 10 cm. and 60 cm.Hg are compared with those computed from the measurements on the individual lines at 0·5 mm. pressure. At the high pressures no lines are resolved, and the computation, which involves only simple. though lengthy, numerical work, is based on the assumption that the widths of the lines vary directly as the pressure.

§ 2. THE EXPERIMENTS

2.1. *The method*

The absorption in ammonia at pressures approaching an atmosphere is so large that it may conveniently be measured using path lengths of the order of a metre. The experiments of Cleeton and Williams (1934) showed that radiation of a wave-length of 1·25 cm. decays to half intensity in a distance of about 80 cm. These experiments were carried out by methods similar to those used in the infra-red, using a large mirror to collimate the radiation; the ammonia

---

* For a narrow line the line-breadth constant equals half the width of the line at half-intensity; it is thus half of what is generally termed the " line-width " in spectroscopy. (See equation (2) below.)

was contained in a cloth bag which could be introduced between oscillator and detector. For accurate measurements, collimation of the radiation by means of hollow wave-guides is much superior, and a cell of rectangular wave-guide one metre long was used in the experiments now to be described. The ends of the cell were sealed with thin mica windows so that it could be evacuated or filled with ammonia at any desired pressure up to about one atmosphere. Waves of the H_{01} type were excited in the wave-guide by a reflection klystron oscillator, and their intensity was measured by means of a vacuum bolometer placed on the far side of the ammonia cell. The absorption due to the ammonia was determined from the reduction in the intensity registered by the bolometer on admitting gas to the cell. The attenuation due to the gas is greater in a wave-guide than in the unbounded medium by a factor

$$\frac{c}{v_g} = \left(1 - \frac{\lambda^2}{\lambda_c^2}\right)^{-\frac{1}{2}},$$

where v_g is the group velocity of the waves in the guide. The critical wave-length λ_c may be calculated from the dimensions of the wave-guide, and hence the attenuation in the guide can be converted to absorption coefficent in the unbounded medium.

2.2. *The apparatus*

Wave-guide windows. Reflections from the windows of the wave-guide cell cause a drop in the power registered by the bolometer; this would not matter if it remained constant during the experiment. Since the attenuation within the cell is small, interference takes place, however, between the reflections from the windows at either end, and the amount of power reflected depends on the phase difference between the two reflections. This phase difference changes when ammonia is admitted to the cell, because of the small increase in the optical path-length, and the change in the net power reflected by the windows causes an error in the measurement of the absorption. This error is greatest when the change in the optical path-length is about a quarter of a wave-length, as is the case at the highest pressures used in these experiments. Owing to the absorption in the ammonia, the total power reflected by the windows cannot be determined accurately by measurements of the standing wave pattern in the guide preceding the cell; it is therefore important to minimize the error by making the reflection coefficient of each window as small as possible. By use of thin mica (0·002 in. thickness was the smallest that would withstand the pressure of the atmosphere) and careful design of the mounting of the window in the guide, the voltage reflection coefficient was generally kept below 0·1. The corresponding error in the measurement of the attenuation should not exceed ±0·1 decibels.

Oscillator. The source of power was a reflection klystron oscillator of a type designed and constructed in this laboratory by Dr. D. Roaf. To cover the range of these experiments (0·63 to 0·92 cm.$^{-1}$), three tubes were used, and the power available was generally over 10 milliwatts.

Attenuating sections. Matched attenuating sections were inserted (*a*) between the oscillator and test cell, and (*b*) between the test cell and the detector. These ensured that the oscillator worked into a load of constant impedance, and that the source and termination were correctly matched into the wave-guide. A variable attenuator which preceded the bolometer served to adjust the power level.

Detector. The power available at the detector (about 100 microwatts) was sufficient to be measured by a vacuum bolometer, whose sensitive element was a piece of fine steel wire. The bolometer formed one arm of a Wheatstone-bridge circuit. Radio-frequency power was estimated from the out-of-balance current due to the increased bolometer resistance. Preliminary experiments showed that the change in resistance was very nearly, but not exactly, proportional to the incident power. A calibration was therefore made by supplying the bolometer with known amounts of D.C. power. Thus the power measurements were more reliable than they would have been if a crystal rectifier had been employed as detector.

Wave-length measurement. Cylindrical cavities employing the H mode of resonance were used for measuring the wave-length. Three wavemeters of this type covered the range 0·63 to 0·92 cm.$^{-1}$.

2.3. *Experimental procedure*

The absorption was measured at regular intervals of about 0·02 cm.$^{-1}$. The oscillator was set to give the desired frequency, and the power registered by the detector while the cell was empty was observed. This was facilitated by an r.f. switch, consisting of a metal strip which could be inserted into the wave-guide to cut off the detector from the radiation; the galvanometer reading corresponding to zero power in the bolometer could thus be observed at frequent intervals without switching off the oscillator. The latter would be a very undesirable procedure because of the time taken to reach a steady output after any adjustment of the potentials applied to the electrodes of the klystron, and because the tube does not return exactly to the same power output.

Ammonia was then admitted to the wave-guide cell, its pressure being measured on a mercury manometer, and the bolometer reading was taken. This procedure was repeated at a number of different pressures, and the bolometer reading with the cell evacuated, and with zero power, again observed. The zero drift was usually about two to three millimetres, and the change in deflection (about thirty centimetres) for the empty cell was of the same order. A small correction was therefore applied assuming the drifts to have taken place uniformly.

§ 3. RESULTS

At each wave-length, a curve of attenuation in the wave-guide against pressure was drawn; two typical curves are shown in figures 1 and 2, where the reduction to absorption in the unbounded medium has been made to facilitate comparison with the results of other workers, and the pressures (cm. Hg) are plotted as abscissae on two scales differing by a factor 10. Figure 1 shows the results of measurements at a wave-length of 1·25 cm. ($\bar{\nu} = 0{\cdot}80$ cm.$^{-1}$), which is close to the centre of gravity of the absorption. The curve rises steeply at low pressures, where the contributions to the absorption from the tails of nearby lines increase rapidly as their widths are increased. Then at high pressures the curve becomes much flatter; this is to be expected, since the intensity at the centre of a pressure-broadened line is independent of the pressure. In this case, the absorption is due to a number of lines, but at sufficiently high pressures, where the widths of the lines are greater than their separation, the attenuation should tend to a constant value.

Figure 2 shows the results at $\lambda = 1{\cdot}47$ cm. ($\bar{\nu} = 0{\cdot}68$ cm.$^{-1}$) which lies on the low-frequency side of most of the strong lines. At pressures of a few cm. Hg, the curve is markedly different from the previous curve, being concave upwards; this corresponds to the fact that the attenuation at a point on the tail of a line should rise with the square of the pressure, as both the number of the molecules per c c. and the width of the line are increasing. As a number of lines are involved, the curve does not rise as steeply as this, the nearby lines giving a constant contribution when their width becomes greater than the difference between their resonant frequencies and the frequency of measurement, while the contributions from the more distant lines are still rising sharply. At the highest pressures the curve becomes flatter, and should ultimately behave like that for $\bar{\nu} = 0{\cdot}80$ cm.$^{-1}$.

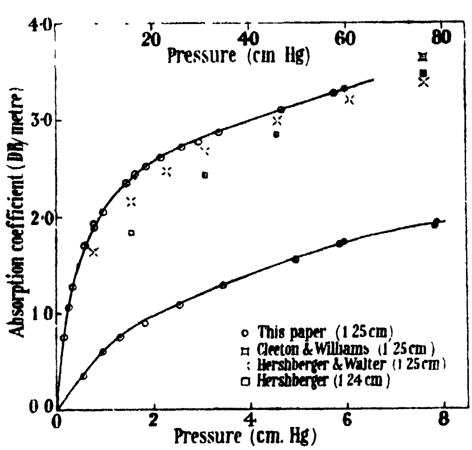


Figure 1. Absorption in ammonia at 1·25 cm. wave-length.

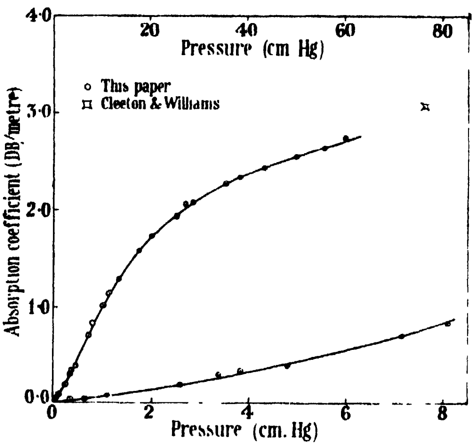


Figure 2. Absorption in ammonia at 1·47 cm. wave-length.

The limiting attenuation will, however, be different since the absorption coefficient is proportional to the square of the frequency.

The results of other experimenters are shown in figures 1 and 2 for comparison. The pioneer work of Cleeton and Williams at atmospheric pressure is represented by a point on each figure; these points appear to be in close agreement with the values predicted from extrapolation of the measurements described in this paper. Recent measurements by Hershberger at wave-lengths of 1·25 cm. and 1·24 cm. are also shown in figure 1, the points being taken from figure 3 of his first paper (Hershberger, 1946) and figure 2 of Walter and Hershberger (1946) respectively. At pressures above 10 cm. Hg, the difference between the absorption coefficients at these two wave-lengths should be less than 3%, as can be seen from figures 4 and 5 of this paper, and as would be expected from the fact that the line-breadth constants at these pressures are 0·1 cm.$^{-1}$ or greater. The large difference between the absorption at these two wave-lengths found by these workers is thus very difficult to understand. As no experimental points are given, nor any estimate of the accuracy, the discrepancy may be attributed to experimental error, and it will be seen that it is of the same order as the difference between Hershberger's values at 1·25 cm. and our curve.

From the graphs of figures 1 and 2, and similar graphs for other wave-lengths, smoothed values for the absorption at a number of selected pressures may be obtained. These absorptions are then reduced to the values appropriate to propagation in an unbounded medium, by multiplying by the ratio of the velocity of light to the group velocity in the guide. The susceptibility of the ammonia gas is so small that it may be neglected in this calculation. The shape of the absorption curves at various pressures may be found by plotting the absorption as a function of the wave-number (figure 3). The pressures chosen were 1, 2, 5, 10, 30 and 60 cm. Hg. The number of points on the curves for the two lowest pressures is inadequate to delineate them accurately, but was sufficient in the preliminary survey of the spectrum to show that a fine structure existed, the separation of whose components must be considerably greater than that suggested by the calculations of Sheng, Barker and Dennison (1941). This fine structure has since been investigated in detail at lower pressures, and these low-pressure curves are not now of particular importance.

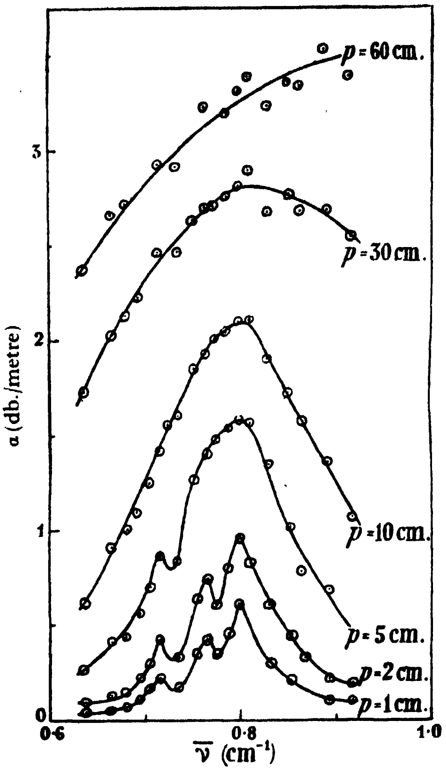


Figure 3. Absorption curves for ammonia at various pressures.

At a pressure of 10 cm.Hg, the lines are sufficiently broad to obscure the fine structure completely, and the absorption curve is quite smooth. The smallness of the scatter of the experimental points is very gratifying, and is well within the possible experimental error. At the high pressures the absorption curves become flatter, and the scatter in the measurements is markedly greater, about ±4% in the curve for 60 cm. pressure. This may be attributed to the small reflections from the windows of the cell, which set up an interference pattern. The introduction of ammonia at 60 cm. pressure changes the optical distance between the windows by more than a quarter wave-length, and the scatter in the points is consistent with the fluctuation of about ±0·1 db. in the power reaching the detector, calculated in the preceding section. The magnitude and sign of the fluctuation change rapidly with wave-length as the cell contains some 60 wave-lengths; thus the scatter in the points on the curve for 60 cm. pressure may be ascribed to this cause, a conclusion that is supported by the smaller scatter at 10 cm. pressure, where the change in the optical path-length is less.

§4. DISCUSSION

As mentioned in the Introduction, the main interest of these measurements at high pressures arises in their use for the study of the collision broadening of spectral lines. The complicated nature of the spectrum of ammonia makes the comparison of theory and experiment considerably more intricate than for a single line, and the shape of the absorption curve at high pressures can only be computed by numerical methods. The data upon which the calculation must depend have been determined in the analysis by the authors of the spectrum at low pressures (Bleaney and Penrose, 1946 b). The data required are:

(1) *The frequencies of the component lines of the fine structure*

These are accurately given by the formula

$$= 0{\cdot}7935 - 0{\cdot}0050_5(J^2+J) + 0{\cdot}0070_4K^2 + 0{\cdot}63 - \{0{\cdot}0050(J^2+J) + 0{\cdot}0070K^2\}^2 \text{ cm.}^{-1}\ (\textit{vacuo}).$$

This formula has received independent confirmation from the experiments of Good (1946).

(2) *The intensities of the lines*

These are given by the theoretical formula

$$\int \frac{\alpha}{\nu^2}\,d\nu = \frac{8\pi^3 N_{JK}}{3ckT}\,|\mu_{JK}|^2, \qquad \ldots\ldots(1)$$

where α = absorption coefficient per cm. of path, $|\mu_{JK}|^2 = \mu^2 \frac{K^2}{J^2+J}$, and N_{JK} is the number of molecules per c.c. occupying the rotational level characterized by the quantum numbers J, K. The authors have shown that the total intensity at a pressure of 4·5 mm. Hg agrees very closely with that calculated from this formula, assuming the value of $1{\cdot}44 \times 10^{-18}$ e.s.u. for the dipole moment μ. This value is confirmed by the most recent measurements of the dielectric constant of ammonia (Van Itterbeek and De Clippeleier, 1946), which yield the value of 1·437 e.s.u.

(3) *The shape of a line at high pressure*

Van Vleck and Weisskopf (1945), and Fröhlich (1946), have shown that the collision-broadening theory of Lorentz is inadequate when the width of the line become comparable with the resonant frequency. Their modification of the theory leads to a structure factor *:

$$F(\nu_0, \nu) = \frac{1}{\pi}\left\{\frac{\Delta\nu}{\Delta\nu^2 + (\nu_0+\nu)^2} + \frac{\Delta\nu}{\Delta\nu^2 + (\nu_0-\nu)^2}\right\}. \qquad \ldots\ldots(2)$$

The absorption coefficient at a frequency ν due to a single line of resonant frequency ν_0 is then given by the expression

$$\alpha = \frac{8\pi^3\nu^2 N_{JK}}{3ckT}\,|\mu_{JK}|^2 F(\nu_0, \nu). \qquad \ldots\ldots(3)$$

---

* Van Vleck and Weisskopf define a "shape factor" which differs from expression (2) by a factor (ν_0/ν). It seems more logical, however, to exclude this factor from an expression for the shape of the line, and to define the "structure factor" so that $\int_0^\infty F d\nu = 1$. The absorption at a particular frequency is then given by multiplying the intensity of the line by the structure factor, the intensity being determined solely by the number of molecules in the two levels between which a transition is taking place and by the probability of such a transition.

(4) *The widths of the lines*

The structure factor $F(\nu_0, \nu)$ involves the line-breadth constant $\Delta\nu$, which for a narrow line equals half the breadth of the line at half intensity. The widths of seventeen lines have been determined by the authors at a pressure of 0·5 mm. Hg (Bleaney and Penrose, 1946 b), the remaining lines not being sufficiently well resolved at this pressure to make an accurate determination of their widths possible. Since the widths vary from line to line, some means of estimating the widths of the unresolved lines is necessary. It has been found that the line-breadth constants of the seventeen lines at a pressure of 0·5 mm. Hg can be expressed by the formula

$$\Delta\bar{\nu}\,(\text{cm.}^{-1}) = 5{\cdot}0 \times 10^{-4}\,\sqrt[3]{K^2/(J^2+J)} \qquad \ldots\ldots(4)$$

For present purposes this formula may be regarded as empirical; its theoretical significance will be discussed in another paper. The deviations of the measured values of the line-breadth constants from those calculated by this formula lie within the experimental error, and (4) may be used for the line-breadth constants of all the lines.

These data are sufficient to calculate the shape of the absorption curve at any pressure, provided some basic assumption is made as to the way in which the line-breadth constant varies with pressure. The simple and obvious assumption is that the line-breadth is directly proportional to the pressure, so long as the latter is not so high that the chance of multiple encounters becomes appreciable. The extension of equation (4) to other pressures is then

$$\Delta\nu\,(\text{cm.}^{-1}) = 1{\cdot}00 \times 10^{-2} p_{\text{cm.}}\,\sqrt[3]{K^2/(J^2+J)}. \qquad \ldots\ldots(4\,a)$$

The two pressures chosen for computation were 60 cm. Hg and 10 cm. Hg. The former is the highest pressure at which experiments were made, while the latter is a suitable intermediate pressure where the scatter in the experimental points is small. The absorption at these two pressures varies only slowly with frequency, and the form of the calculated curves can be determined by computation at a reasonably small number of frequencies. The form of equation (3) shows that it is preferable to plot α/ν^2 as a function of the frequency or $\alpha/\bar{\nu}^2$ against the wave-number $\bar{\nu}$, rather than the actual absorption coefficient α, as the presence of the ν^2 term in the latter causes the maximum absorption to shift to higher frequencies when the line becomes broad. This shift is not due to displacement of the resonant frequency and tends to obscure any such effect.

The experimental values of $\alpha/\bar{\nu}^2$ for a pressure of 10 cm. Hg are shown plotted against $\bar{\nu}$ in figure 4. The curve computed from the measurements at 0·5 mm. pressure is drawn as a full line, constructed from points calculated at intervals of 0·05 cm.$^{-1}$ between 0·6 and 1·0 cm.$^{-1}$. It will be seen that the experimental points lie very well on the calculated curve, the deviations being within the estimated experimental error. The following conclusions may therefore be drawn:—

(*a*) There is no doubt that the widths of the lines vary accurately as the first power of the pressure. If the widths varied with the square root of the pressure, as is suggested by some experiments (see Elsasser, 1942), the calculated absorptions would be too small by a factor of $\sqrt{200}$, since the widths have been determined at a pressure of 0·5 mm. Hg and are here extrapolated to a pressure of 10 cm. Hg.

(*b*) There is no significant shift in the resonant frequency of the lines at this pressure.

In figure 5 the experimental values of $\alpha/\bar{\nu}^2$ at a pressure of 60 cm. Hg are shown, together with the calculated curve. There is now an appreciable discrepancy between the measured and the calculated values, the experimental points indicating that the absorption is materially greater than expected at the low-frequency end. The maximum has moved well away from 0·78 cm^{-1} to lower frequencies; no such shift would be predicted by the structure factor. The experimental values do not extend to sufficiently low frequencies to determine the position of the maximum, but in any case the discrepancy between them and the calculated curve is considerably greater than the experimental error. In any attempt to explain the discrepancy, the following possibilities must be considered:

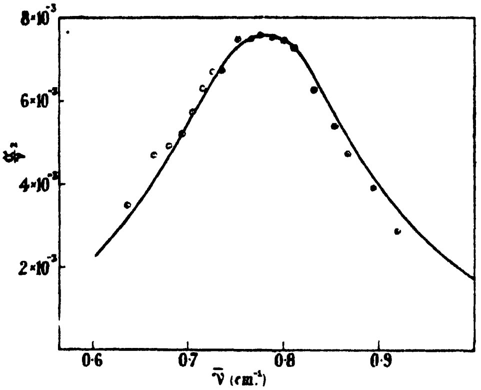


Figure 4. Curve of (absorption coefficient per cm. against wave-number²) at a pressure of 10 cm. Hg.

○ Experimental points.
—— Calculated curve.

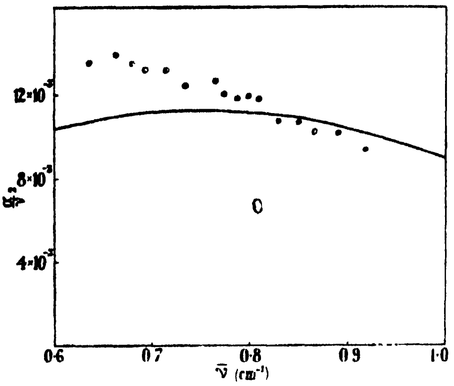


Figure 5. Curve of (absorption coefficient per cm. against wave-number²) at a pressure of 60 cm. Hg.

○ Experimental points.
—— Calculated curve.

(*a*) At high pressures the resonant frequencies of the lines are displaced towards lower frequencies, due to interaction with neighbouring molecules. Since ammonia is a strong dipole, the most important interaction should be a Stark splitting due to the electric field. Polder (1942) has pointed out that only a quadratic effect would be expected in weak fields, the two energy levels of the inversion doublet being given by the formula of Penney (1931),

$$W_{1,2} = \pm\sqrt{\left(\frac{\Delta}{2}\right)^2 + (\mu' E)^2},$$

where

$$\mu' = \mu\frac{MK}{J(J+1)}.$$

Expansion in ascending powers of $(\mu' E)^2$, assuming the latter to be small compared with $(\Delta/2)^2$, gives for the energy difference between the two levels*

$$W_1 - W_2 = \Delta + \tfrac{1}{2}\frac{(\mu' E)^2}{\Delta}. \qquad \ldots\ldots(5)$$

This formula indicates that a displacement towards higher frequencies should occur; if the mean value of E is taken as μ/d^3, where d is the mean distance between the molecules of the gas, the displacement would amount to a few hundredths of a cm.$^{-1}$ at a pressure of 60 cm. Hg. The Stark-effect shift is thus not only of the wrong sign, but also of the wrong magnitude to explain the observed shift.

(*b*) At high pressures the widths of the lines are not directly proportional to the pressure.

At 60 cm. pressure the mean distance between the molecules is about 37 A., while for the broadest lines a collision occurs when the molecules are some 14 A. apart; the chance of multiple encounters is thus appreciable. It is therefore difficult to estimate the sign or magnitude of the deviations from the linear law, but the deviations should be greatest for the widest lines, since these are due to molecules with the largest collision cross-section. These lines lie in the main towards the high-frequency side (the strongest lines, for which $J = K$, all lie at wave-numbers $>0{\cdot}79$ cm.$^{-1}$); and they are all so broad ($\Delta\bar{\nu} \approx 0{\cdot}5$ to $0{\cdot}6$ cm.$^{-1}$) that the absorption due to them is practically constant over the range of wave-numbers of these measurements. The main effect of a change of the width is therefore to shift the curve bodily up or down without material change of shape. On the other hand, to explain the greater absorption observed on the low-frequency side it would be necessary to postulate that the widths of the narrower lines deviate most from the linear law, the widths increasing less rapidly than the first power of the pressure. It thus appears unlikely that the shape of the observed absorption curve could be explained by a simple deviation from the ($\Delta\bar{\nu} \propto p$) law, though the evidence is not sufficient to rule it out.

(*c*) An appreciable amount of radiation is absorbed by a molecule "during a collision".

When the mean distance between the molecules is comparable with the collision diameter for this absorption, a molecule will spend an appreciable fraction of the total time within regions where the electric field is of the order of 10^6 volts/cm. or more. During this time the energy levels of the molecule must be greatly distorted; there is no simple method of estimating the effect on the absorption, but at high pressures, where multiple encounters are the rule, one might expect to obtain a Debye curve similar to those observed in the case of liquids.

---

* *Note added in proof.* The second-order Stark effect in ammonia has been observed by Coles and Good (1946), who find that the splitting can be represented by the formula

$$\delta\bar{\nu}\ (\text{cm.}^{-1}) = 1{\cdot}5 \times 10^{-4}\,[MK/(J^2+J)]^2 E^2.$$

This is in good agreement with (5), which, on insertion of the numerical values of Δ and μ, gives

$$\delta\bar{\nu}\ (\text{cm.}^{-1}) = 1{\cdot}3 \times 10^{-4}\,[MK/(J^2+J)]^2 E^2.$$

§ 5. CONCLUSION

The inferences which can be drawn from these measurements may be summed up as follows:—

(*a*) The widths of the lines vary linearly with the pressure between 0·5 mm. and 10 cm. Hg.

(*b*) There is no appreciable shift in the resonant frequencies of the lines at pressures up to 10 cm. Hg.

(*c*) The shape of the absorption curve at 10 cm. pressure conforms to that calculated from the structure factor of Van Vleck and Weisskopf, but the measurements do not extend far enough in frequency to provide an adequate test of this factor.

(*d*) At a pressure of 60 cm. Hg, the interactions between the molecules are so large that the absorption curve is appreciably distorted.

It is obvious that considerable extension of the measurements described in this paper is desirable, and has already begun. The frequency range is to be increased so as to include substantially the whole of the extent of the absorption, and to afford an experimental test of the structure factor. The distortion of the absorption curve at 60 cm. Hg should then be more obvious, and will be examined at still higher pressures. Preliminary measurements indicate that the shift in the maximum of the curve of $\alpha/\bar{\nu}^2$ against $\bar{\nu}$ will become even more pronounced at higher pressures. This is indicated by the fact that the curves of absorption against pressure for $\bar{\nu}=0{\cdot}7$ cm.$^{-1}$ begin to turn upwards again at about 90 cm. Hg instead of flattening off, while similar curves for $\bar{\nu}=0{\cdot}9$ cm.$^{-1}$ do not. This could not be the case if the conclusions (*a*) and (*b*) above were still valid at these pressures.

ACKNOWLEDGMENT

The work described in this paper was carried out on behalf of the Director of Physical Research, Admiralty, and the authors wish to record their thanks for permission to publish.

REFERENCES

Bleaney and Penrose, 1946 a. *Nature, Lond.*, **157**, 339.
Bleaney and Penrose, 1946 b. *Proc. Roy. Soc.*, A, **189**, 358.
Cleeton and Williams, 1934. *Phys. Rev.*, **45**, 234.
Coles and Good, 1946. *Phys. Rev.*, **70**, 979.
Elsasser, 1942. *Harvard Meterological Studies*, No. 6, pp. 45–48.
Fröhlich, 1946. *Nature, Lond.*, **157**, 478.
Good, 1946. *Phys. Rev.*, **70**, 213.
Hershberger, 1946. *J. Appl. Phys.*, **17**, 495.
Penney, 1931. *Phil. Mag.*, **11**, 602.
Polder, 1942. *Physica*, **9**, 908.
Sheng, Barker and Dennison, 1941. *Phys. Rev.*, **60**, 786.
Van Itterbeek and De Clippeleier, 1946. *Physica*, **12**, 97.
Van Vleck and Weisskopf, 1945. *Rev. Mod. Phys.*, **17**, 227.
Walter and Hershberger, 1946. *J. Appl. Phys.*, **17**, 814.

THE FUNDAMENTAL CONCEPTS CONCERNING SURFACE TENSION AND CAPILLARITY

By R. C. BROWN,
University College, London

MS. received 8 *November* 1946; *lecture delivered* 6 *December* 1946

ABSTRACT. It cannot be denied that our elementary discussions of the principal phenomena connected with surface tension and capillarity are often vague and unsatisfying. The simple statics of the systems dealt with are frequently obscured and circumvented by the introduction of the concept of free surface energy to replace surface tension. The lecture is an attempt to clarify some of the points which arise from this failure to come to grips with fundamentals.

On the basis of the usual idea of cohesion between molecules it can be shown, for instance, that, in contradiction to what is often contended, it is not necessary to deny the reality of surface tension during the course of an explanation of the common phenomena which were, at one time, regarded as providing evidence of its existence. It is possible, also, to gain a less abstract conception of the distinction between free and total surface energy than that provided by a purely thermodynamical discussion.

The customary assumption that in a system containing a solid/liquid interface the surface energy of the solid plays an identical rôle with that of the liquid is criticized, and the conception of surface energy as the work done during the rupture of a column of material is examined. Capillary elevation is regarded as a consequence of negative surface tension in the liquid at the solid/liquid interface, and the usual expression for the capillary rise is derived from this idea.

§ 1. THE EXISTENCE OF SURFACE TENSION

It has become customary in recent years for authors of text-books to deny the existence of surface tension, especially (and somewhat incongruously) when they are presenting an elementary version of Laplace's theory. It is frequently made to appear that the theory explains *away* surface tension instead of accounting for it in a physical manner.

Naturally enough, however, the non-existence of surface tension is not emphasized on those pages which deal with its experimental determination.

Examples of the belief in the non-existence of surface tension are numerous. Thus Newman and Searle (*The General Properties of Matter*) write: "It should be realized, however, that the term 'surface tension' is misleading by reason of its suggestion that there is a real stretching force tangential to the surface of a liquid" (page 163). Champion and Davy (*Properties of Matter*) refer to surface tension as a "useful fiction" (page 99) while Adam (*The Physics and Chemistry of Surfaces*) regards it as a "purely mathematical device" (page 4), and maintains that ".... surface tension does not exist as a physical reality, and is only the mathematical equivalent of the free surface energy" (page 5).

It is clear that we must ask how it comes about that when a soap film supports a weight it should be believed to do so by virtue of a "useful fiction" or a "mathematical device", while a sonometer wire or a clothes line is credited with no such mystical power but is supposed to be in a perfectly real state of tension.

As already mentioned, denials of the reality of surface tension appear to arise during the elementary consideration of the hypothesis which for brevity we shall call "Laplace's theory". It is explained that if we postulate the existence of cohesive forces between molecules of a liquid, then each molecule which is situated in or near the surface is acted upon by a force directed inwards and normal to the surface. It is therefore necessary to do work against this force in order to take a molecule from the interior to the surface, and consequently the surface molecules possess greater energy than those inside the liquid. Therefore, in accordance with the principle that every system moves towards a state of minimum potential energy, if free to do so, the surface of a liquid shows a tendency to contract. It is at this stage that we usually come across a statement to the effect that surface *energy* is therefore the reality and that surface tension is a fiction which is "mathematically equivalent" to surface energy.

There are two points to be mentioned in connection with this argument. In the first place, as with all natural phenomena, our belief in the existence of a tension in the tangent plane of a liquid surface must, of course, be based on experimental evidence, and this is surely convincing. We know that in order to maintain a soap film in equilibrium it must be acted upon by an external force parallel to its surfaces. Figure 1 represents a section of film stretched between wires of circular section A and B. If B is fixed and A is free to move, then in order to maintain equilibrium A must be acted upon by an external force *F*. Therefore, if we take any section of the film at, say C, we conclude that for the equilibrium of AC, the film to the right of C must act on the portion to the left with a force equal and opposite to *F*. This force is found to be proportional to the length of the film perpendicular to the plane of the paper and not to the area of cross-section, which suggests that it has its seat in the surface. In fact, it could not act in the interior of the liquid because this would imply that the pressure in the interior of the film was lower than that in the surrounding atmosphere, and also that a change of shape of the film due to an extension would be opposed by a force acting in the interior, which is contrary to experience, which shows that fluid matter cannot sustain a shear.

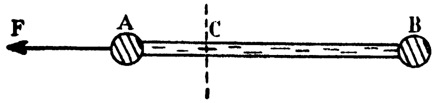


Figure 1.

Secondly, it is erroneous to conclude that surface tension is non-existent merely because a certain elementary development of the postulate of molecular cohesive forces of limited range leads in the first place to an explanation of surface energy. On the contrary, if surface energy exists, it must be necessary to perform work in order to extend a surface, and therefore, when a plane soap film is stretched, there must exist a force parallel to its surface which opposes the stretching and necessitates the application of an external force and the performance of work. It is true that this argument may be regarded as only an indirect way of showing that cohesive forces give rise to surface tension in liquids, but that is a very different matter from denying its existence.

When a liquid surface is extended, the external work is not performed *directly* on the molecules which are brought into the surface from the interior. The system is like a simple machine, the load being the force necessary to transfer the molecules to the surface and the effort being the external force.

In the soap-film example which has just been discussed, the external force balances the surface tension directly, while in the case of the expansion of a bubble or drop by the injection of air or liquid the external force does work against the excess pressure in the interior. We know by a simple argument in statics that this excess pressure implies a tension in the surface. In whatever way the surface of a liquid is extended, we find that the necessity for performing work, which is forecast by Laplace's theory, can be traced to the existence of an opposing force acting tangentially to the surface. The fact is, therefore, that any theory which accounts for surface energy in liquids must, of necessity, also account for surface tension. Free energy per unit area and tension per unit length in a liquid surface are physically equivalent (not merely "mathematically equivalent" as is so often stated). Bearing this in mind, we conclude that it must be possible to develop the original postulates of the theory in such a way as to lead to a more direct explanation of surface tension than is to be found in the usual method of treatment. A simple argument which achieves this is presented in the next paragraph.

Instead of leaving the discussion at the stage where it is shown that the inward cohesive forces acting on the surface layer cause the surface to assume the smallest area compatible with the action of other factors, we should note that when this minimum area has been reached the molecules in the surface are still attracted inwards and possess potential energy with respect to those in the interior. Objects do not remain in positions of high potential energy unless the force by virtue of which they possess that energy is balanced by some other equal and opposite force. Suppose that the surface has reached its minimum area and that the concentration of molecules near the surface is the same as it is in the interior. This condition cannot persist because there is no force to balance the inward attraction experienced by the surface molecules. These will therefore continue to pass into the interior until the reduced concentration in the surface sets up a pressure gradient opposing the inward attraction. When equilibrium is reached, therefore, there are fewer molecules per unit volume in the surface than in the interior and we know from our experience of the effect of pressure on the volume of a liquid that a smaller density means a smaller pressure. Thus the surface layer is in a state of tension compared with the condition of the interior.

It will be noticed that the *fluid* character of the substance is essential to this argument. There are no shearing forces within the liquid itself to oppose the inward flow of molecules until the lack of molecules near the surface establishes an opposing force. The lower molecular concentration in the surface is an example of the general principle that molecules distribute themselves with smaller concentrations at places of higher potential energy.

When equilibrium is established, the potential energy which a molecule would lose by moving from the surface to the interior is equal to the potential energy of strain which the surface layer would gain by this transfer (cf. a small downward displacement of a body supported by a helical spring). In other words no molecules have any resultant force acting upon them, otherwise they are not in equilibrium.

A deeper analysis of the equilibrium between the surface and the interior has been worked out by Bakker (Vol. VI, Wien-Harms' *Handbuch der Experimentalphysik*) and it seems worth while to give a brief outline of this treatment.

Imagine a small plane surface of area Δs drawn anywhere in a liquid and let F be the force with which the liquid on one side of the plane attracts that on the other side. Then the *cohesion*, K, is given by

$$K = F/\Delta s.$$

Cohesion is also known as *intrinsic pressure*, and is an extremely large pressure of the order of 10000 atmospheres or more.

In addition to K there is, in general, another force which also presses together the two portions of liquid separated by Δs. This is, of course, what we ordinarily know as the pressure in the liquid, a quantity which, unlike K, is measurable directly by the force on a piston, one side of which is exposed to the liquid. It is produced by agents external to the liquid itself such as gravity, the atmosphere, artificial compression in a cylinder, or simply by the vapour pressure if the liquid partially fills a closed vessel.

If the value of the pressure is p at any point where the cohesion is K, then the *total pressure*, P, at this point is given by

$$P = K + p.$$

Thus the force per unit area with which the liquid on one side of an imaginary element of area presses against that on the other side is made up of the force due to cohesion and the force due to external causes. In other words, the condition of compression of the liquid is the same as it would be if no cohesion existed but a pressure equal to K were applied externally in addition to p. It is necessary to use two different terms to distinguish between P and p, and we shall adopt *total pressure* for the former and simply *pressure* for the latter.

The crux of the argument which demonstrates that cohesion gives rise to surface tension is the fact that the *total pressure* at a point in a fluid has the same value in all directions. Although we are accustomed to regarding this as being a property of the ordinary pressure, p, yet this is actually a particular case which is true only at points where K has the same value in all directions.

In the elementary proof that the pressure at a point in a fluid is independent of the orientation of the plane across which it is measured we consider the action of the external forces due to p upon the three faces of a prismatic element of fluid. In fact, however, the total force per unit area on any face is not p but $K+p$, i.e. P. It is this force which would have to be applied to the faces of the element (by solid pistons for example) in order to maintain its original condition if it were isolated from the body of the liquid. Thus the property of being independent of direction is fundamental to P rather than to p. We shall see that at any point in the surface layer K is a function of direction so that p is also dependent upon direction in this region and it is this which gives rise to surface tension.

Suppose that the attractive force between two small elements of the liquid becomes negligible when the distance between them exceeds a certain small value c. Then if we consider a quantity of liquid in the form of a column of uniform cross section Δs it is clear that the force of attraction between the two parts into which any cross section divides the column is independent of the position of the section provided that its distance from either end of the column is not less than c. Therefore a surface layer in which K is less than it is in the interior extends inwards from each of the end faces of the column. If the substance is supposed to be inextensible the depth of this layer is equal to c.

Figure 2 represents a portion of the surface layer of a liquid, that is to say, the region in which a particle of liquid experiences a net inward attraction. Let a small area Δs be drawn in the surface layer parallel to the surface.

Then, if at the level of Δs the ordinary fluid pressure taken perpendicular to Δs is p, and if K_n is the cohesion taken in the same direction, we can write

$$P = K_n + p, \qquad \ldots\ldots(1)$$

where P is the total pressure at Δs.

Since K_n is the attractive force per unit area between the two portions into which the cross-section Δs divides the vertical column indicated in figure 2, the value of K_n must be zero at the outer boundary of the surface layer (if we suppose that the vapour exerts no attractive force on the liquid) and it increases as Δs moves into the liquid, until at the inner boundary of the layer its value is equal to the cohesion in the interior of the liquid.

We shall ignore the effect of gravity upon conditions in the surface layer, and regard p as being entirely due to the atmosphere or vapour above the liquid surface. By considering the equilibrium of the column lying between Δs and the outer boundary of the surface in figure 2 we observe that, in the liquid, the pressure p taken across an element of area parallel to the surface is everywhere equal to the pressure in the medium above the surface.

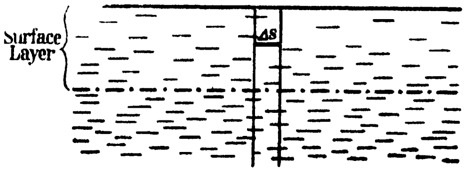


Figure 2.

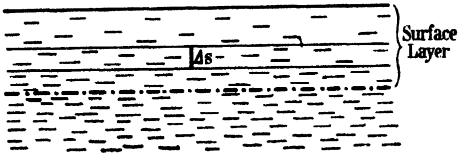


Figure 3.

If we now take Δs at the same level in the surface layer as previously, but perpendicular to the surface (figure 3), we have

$$P = K_s + p_s \qquad \ldots\ldots(2)$$

where K_s and p_s are respectively the cohesion and pressure taken across a small area perpendicular to the surface at this particular depth. Since P is the same in equations (1) and (2), we have

$$p - p_s = K_s - K_n.$$

Now, when Δs is perpendicular to the surface (figure 3), the columns of attracting liquid extend to a distance greater than c on each side of it, whereas when it is parallel to the surface (figure 2) the column on the side nearer the surface is shorter than c, and the attractive force across Δs is therefore less. Thus it follows that for any given point in the surface layer

$$K_s > K_n$$

so that

$$p_s < p,$$

which implies a state of tension parallel to the surface.

The expression for the surface tension may be derived as follows:—In figure 4 some liquid is contained between three vertical walls of a closed vessel A and a movable partition B attached to a rod which passes through the wall of A. It is convenient to suppose that the angle of contact between the liquid and B is 90°, and this may be done by choosing the appropriate solid without affecting the value of the surface tension of the liquid. If the effect of gravity is left out of consideration, the pressure is uniform throughout the interior of the liquid and is equal to that in the gas and/or vapour confined in A. At all points in the surface layer, however, the pressure p_s across an area perpendicular to the surface is less than that in the vapour. Apart from the effect of the surface layer, the partition B would experience no resultant force, but if the layer has a thickness t and unit width perpendicular to the plane of the paper, the force on B towards the left is equal to

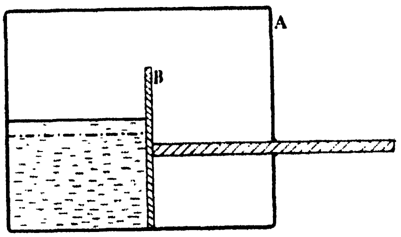


Figure 4.

$$(p-p_s)t,$$

where p_s now represents the *mean* pressure in the surface layer in the direction parallel to the surface. The surface tension (γ), as commonly understood, is equal to the external force which must be applied to the rod in order to keep B in equilibrium and is therefore given by

$$\gamma=(p-p_s)t. \qquad \ldots\ldots(3)$$

At first sight, equation (3) appears to suggest that the pressure (p) in the liquid and surrounding gaseous phase has a direct mechanical effect on the magnitude of γ. This is not the case, however, because an increase of p causes an equal increase to occur in P and therefore also in p_s (equations (1) and (2)). Nevertheless it is worth noting that the *actual* tension in the surface is not γ but $-p_s t$, to which we may give the symbol σ.

As already mentioned, the belief in the non-existence of surface tension appears to arise from the misconception that cohesion in a liquid explains only surface energy, and it is hoped that the foregoing demonstrates that this is, in fact, a misconception. At the same time the author does not admit that surface tension owes its existence to any *theory*, but rather to the simple statics of, say, a plane soap film or bubble.

In the course of a very brief reference to Bakker's treatment, Adam (*The Physics and Chemistry of Surfaces*, p. 4) admits that it is "mathematically correct" but makes the following criticism:—"It is certain now, however, that the surface layer is ordinarily only a few molecules thick, so that the conception of pressure and its variation parallel to the surface becomes rather intangible in terms of molecules." The answer to this is, surely, that we do not regard the surface layer as a collection of stationary molecules but as a *region* within which the molecules experience resultant forces which are absent in the interior. Even if the region is very thin, there is within it an infinite number of possible levels which can be occupied by the molecules instantaneously during their thermal agitation. A different value of P is associated with each level, and over a sufficiently long period of time each level will contain the centres of a definite average number of molecules.

§2. A MECHANICAL ANALOGY

A mechanical model which represents the mechanism of surface tension can be devised as follows. In figure 5, ABCD is a continuous length of cord. The portions AB and CD rest on smooth horizontal shelves at different levels, and pulleys are provided at B and C. The portion AB represents a collection of molecules in the interior of a liquid while CD represents molecules in the surface. We can regard CD as a kind of average level between the inner and outer boundaries of the surface layer of the liquid. The path which the molecules take in moving from the interior to the surface is represented by BC. The molecular attraction against which the molecules of the liquid must move in order to reach the surface is replaced in the model by the earth's gravitational field acting on every element of the cord between B and C, the total force being the weight of BC. In order to maintain the system in equilibrium it is necessary to apply at D an external force (γ) equal to the tension in CD which, in turn, is equal to the weight of BC. If the cord is inextensible and D is caused to move to the right, the work done by γ appears as the gravitational potential energy gained by the length of cord which is raised to the higher level. If the cord is extensible, however, a given amount of external work performed by γ raises a smaller weight of cord than otherwise and the remainder of the work appears as energy of strain because the portion of cord which has been raised has also been extended. If the elastic and thermal properties of the cord are such that it cools when stretched adiabatically it will be necessary to supply heat when the cord is drawn up to the higher level isothermally, so that its behaviour is comparable to that of a liquid surface in this respect also.

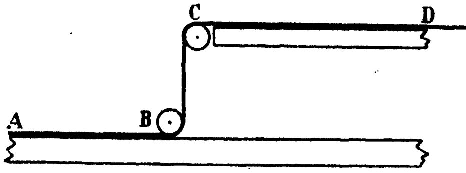


Figure 5.

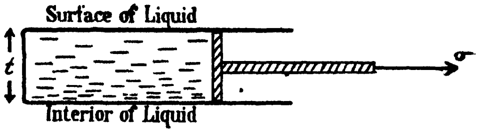


Figure 6.

§3. TOTAL AND FREE SURFACE ENERGY

The well known relation between total surface energy (E) and free or available surface energy (γ), viz.

$$E=\gamma-T\frac{\partial\gamma}{\partial T},$$

is derived from the laws of thermodynamics which are independent of the nature of the working substance and of the processes involved. The equation would apply to any system exerting an external force γ which is a function of temperature. Consequently it frequently happens that little more is said about this aspect of the properties of liquid surfaces than that $\partial\gamma/\partial T$ is negative for the great majority of substances, so that when a surface is extended isothermally by unit area it is necessary to impart to the liquid an amount of heat (viz. $-T\partial\gamma/\partial T$). It is quite possible, however, to gain an idea of the mechanism of surface tension which the equation reflects by considering a "simplified" liquid. Imagine that the surface layer of a liquid is contained in a vessel as shown in figure 6 and that its width perpendicular to the plane of the paper is unity. We can suppose that the lower wall of the vessel separates the surface layer from the interior but does not modify the cohesive forces acting across it. In order to maintain the piston in equilibrium it is necessary to act upon it with an external force directed outwards and numerically equal to σ $(=-p_s t)$, the total tension in the surface layer. This is the *total* external force necessary to keep the surface in equilibrium. In the case of a plane soap film, σ is the resultant of an applied force γ acting outwards and the force pt due to the atmosphere acting inwards (equation (3)).

When the piston is pulled out reversibly through an infinitesimal distance, the pressure is slightly decreased everywhere, and if a hole were now made in the lower wall of the vessel a small amount of liquid from the interior would flow into the surface layer and restore the pressure to its original value. This is a useful way of picturing the mechanism of surface extension. The total external work performed during a finite extension can be regarded as being the sum of all the small amounts of work performed during a series of steps like the one just described, the force σ remaining constant for successive steps because liquid is admitted after every small extension.

Now, since

$$\sigma = p_s t = (K_s - P)t, \qquad \ldots\ldots(4)$$

it is evident that we can regard the force σ which the surface exerts on the piston as being the resultant of a force due to cohesion, $K_s t$, and an opposite force Pt tending to expand the surface. In order to extend the surface by a small area, da, it is necessary, therefore, to do a total amount of work $K_s t\,da$ against cohesion. Of this total amount a quantity of work σda is performed by the external force which balances σ, leaving, by equation (4), $Pt\,da$ units of work to be performed by P against the cohesion. Thus the total pressure, P, assists the extension of the surface by its tendency to force the molecules apart. In performing this work the molecules will cool in the same way as an ideal gas cools when it does work and, for an isothermal extension of the surface, an amount of heat equivalent to $Pt\,da$ units of work must be given to the system.

Thus for unit area of surface we can write:

Total surface energy $= K_s t$.
Free surface energy $= \sigma$.
Superficial "latent heat" $= Pt$.

Suppose that we regard P as being due to the thermal agitation of the molecules and assume that it is proportional to the absolute temperature when the separation of the molecules remains constant, i.e. when the layer contained in the vessel in figure 6 is heated at constant volume. This assumption is in accordance with the ideas underlying an equation of state such as that of Van der

Waals in which the total pressure $(p+a/v^2)$ is equated to $RT/(v-b)$. On this assumption

$$\left(\frac{\partial P}{\partial T}\right)_v=\frac{P}{T},$$

and the superficial latent heat, Pt, can be written as

$$T\left(\frac{\partial P}{\partial T}\right)_v t.$$

We can, furthermore, reasonably suppose that at constant volume, since the separation of the molecules is constant, the value of K_s is independent of temperature, so that by differentiating equation (4) we obtain

$$\left(\frac{\partial P}{\partial T}\right)_v t=\left(\frac{\partial K_s}{\partial T}\right)_v t-\left(\frac{\partial \sigma}{\partial T}\right)_v$$

$$=-\left(\frac{\partial \sigma}{\partial T}\right)_v.$$

The superficial latent heat therefore becomes

$$-T\left(\frac{\partial \sigma}{\partial T}\right)_v,$$

and we can replace equation (4), i.e.

$$K_s t=\sigma+Pt,$$

by the energy equation

$$E=\sigma-T\left(\frac{\partial \sigma}{\partial T}\right)_v,$$

where the symbol E replaces $K_s t$ to represent the total surface energy per unit area.

Thus, using very simple ideas, it is possible to establish an equation which is similar to the equation of free energy. The two are not identical because as we have seen, σ is not the same quantity as the surface tension γ, their difference being equal to the product of the pressure p by the thickness of the surface layer t. The two quantities become identical only when the pressure is zero. Even at atmospheric pressure, however, the difference between them is very small, being of the order of $10^6 \times 10^{-7}$ or 0·1 dyne cm.$^{-1}$ if the thickness of the surface layer is taken to be of the order of a few molecular diameters. Under ordinary conditions, therefore, we can feel justified in regarding the relationship between the free and total energies of a liquid surface as being due almost entirely to the mechanism upon which we have based the derivation of the last equation.

One point of interest in this connection is that we have identified the cohesive force $K_s t$ with the total surface energy, and have assumed that K_s remains constant when the temperature of the surface layer is changed at constant volume. In actual fact the total surface energy, given by $\gamma-T\left(\frac{\partial \gamma}{\partial T}\right)$, does vary with temperature, and two observations can be made concerning this discrepancy. In the first place, in the thermodynamical derivation of this expression for the total energy, the volume of the surface layer is *not* supposed to be constant as the temperature is varied—nor, indeed, is it constant in reality. In these circumstances the cohesion K_s will vary with temperature on account of the changing average distance between the molecules. Secondly, the increased concentration of molecules in the vapour following a rise of temperature will, in practice, diminish the surface tension by reducing the inward attractive force acting on the surface molecules. This last effect, which eventually reduces γ to zero at the critical temperature, is not taken into account in the above discussion, where the only influence of temperature upon σ is supposed to be that which is due to the variation of P with temperature at constant volume. Nevertheless, the foregoing ideas are supported by the experimental fact that at ordinary temperatures, when the effect of the vapour is small, the total surface energy actually varies very little with the temperature, and in some cases is remarkably constant. One example is mercury, as might be expected on account of its low vapour pressure and comparatively low expansion coefficient. It will be realized that constant total surface energy implies a constant value of $d\gamma/dT$, i.e. a linear relationship between surface tension and temperature.

§ 4. THE SURFACE ENERGY OF SOLIDS

The elementary argument which shows that cohesion between molecules leads to surface energy is, as usually presented, equally applicable to both liquids and solids. In both cases the necessary condition is present, namely, that surface molecules are attracted towards the interior. Or, to put the matter in another way, it is necessary to perform mechanical work in order to break a column of cohering matter so as to form two fresh surfaces, whether the substance be solid or liquid. It is frequently stated therefore that, since the step from surface energy to surface tension is "purely mathematical", solids as well as liquids must possess surface tension. It is, however, worth while questioning this argument, if for no other purpose than to discover how it comes about that, while the surface tension of a liquid can be directly determined in a variety of ways commonly applicable to the measurement of a force, the surface tension of a solid is somewhat elusive.

The important point in this connection is, of course, that the spontaneous contraction of a liquid film, which can be prevented only by the application of an external force equal and opposite to the surface tension, has no counterpart in the case of a solid sheet. The essence of the process is the freedom with which the molecules of a liquid can pass from the surface to the interior, that is to say, the inability of a liquid to sustain a shearing stress and to oppose permanently the attempt on the part of the surface tension to change its shape. In a solid the cohesive forces acting inwards on the surface molecules are balanced by elastic reactions set up within the solid itself so that it is unnecessary to apply an external force in order to prevent the contraction of the surface.

It cannot be said, therefore, that a solid possesses surface tension in the sense in which the term is applied to liquids. Many authors, nevertheless, use the terms "surface energy" and "surface tension" indiscriminately in connection with solids—frequently having denied the existence of surface tension elsewhere in their writings.

A good deal of light can be thrown on the nature of surface energy and the distinction between the properties of solids and liquids by examining the fundamental process of the creation of new surfaces by the rupture of a column of matter.

In the first place let it be assumed that the column is composed of extensible material. Supposing that the molecules are in a static condition and are all equally spaced, then in order to effect a rupture it is necessary to create at one cross section an axial separation of molecules equal to c, the maximum molecular separation for which finite cohesion exists. Theoretically this can be brought about by the application of equal and opposite external forces to the molecules lying immediately on each side of a given transverse section. If the forces are applied elsewhere (e.g. at the free ends of the original column) some of the work which they perform will create a state of strain in the material and it would be necessary to subtract the energy corresponding to this in order to arrive at the amount of work required to produce the new surfaces.

In figure 7, AB and CD represent two contiguous layers of molecules to which external forces are applied. As the separation of the layers proceeds, the value of the cohesion K_n parallel to the axis of the column, i.e. perpendicular to the surfaces about to be formed, diminishes not only between AB and CD but also across a plane such as EF situated near to CD because there are now fewer molecules within the distance c from this plane on the left-hand side. This effect extends only for a limited distance on each side of AB and CD since the range of molecular attraction is limited.

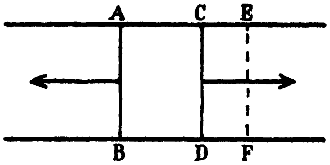


Figure 7.

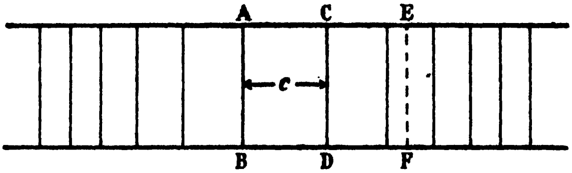


Figure 8.

It is, of course, quite possible that the rupture of a column may not be a completely reversible process. At a certain stage during the stretching, the molecules in the regions which will eventually form the surface layers may become unstable and acquire accelerations which cannot be controlled by adjusting the magnitude of the external forces. The final static state of the substance when rupture is just complete can, however, be discussed without reference to this possibility. In figure 8 the distance between AB and CD has reached the value c. Further separation does not require the expenditure of work, neither does it change the conditions at a plane such as EF. The two parts of the column are now furnished with stable surface layers. If we suppose that the external pressure is zero, then K_n and P (the total pressure) must be equal to each other at every transverse section. Therefore, in the surface layers, e.g. across EF, where K_n is small, P is less than it is in the interior and this implies a greater molecular separation. Thus the condition is as shown diagrammatically in figure 8 where the distances between the transverse lines represent distances between molecules in the direction parallel to the axis of the columns.

The moving apart of the molecules which has occurred as a result of the rupture has, of course, been assisted by the action of P, and the work done by this repulsion must be supplied in the form of heat if the process is to be isothermal. It would therefore appear that when we calculate (as in Laplace's theory) the work done against molecular attraction during the rupture of a column we are in fact finding an expression for the *total* rather than the *free* surface energy, inasmuch as the assistance given by P is ignored.

If we are dealing with a liquid column contained in a tube—for simplicity the angle of contact can be supposed to be 90°—then, as already explained, the diminshed value of K_n in the surface layer sets up a tension parallel to the surface which is capable of doing external work. Superficial latent heat is also accounted for because during a contraction of the surface the closer approach of the molecules to each other as they leave the surface and enter the interior is accompanied by the emission of heat which is the thermal equivalent of the work done during this process by the cohesion between them.

Solid surfaces, however, show no tendency to contract and require no external force to keep them in equilibrium. Nevertheless there is a condition of strain in the surface layer such as that depicted in figure 8, and if this is partially destroyed by, say, allowing a liquid to spread over the solid, thus increasing the value of K_n in the surface layer, some of the heat energy which was required for the establishment of the state of strain will be liberated. The question as to whether mechanical energy also would be derivable *directly* from the disappearance of the strain in the surface depends upon the character of the original process of rupture. If this were completely reversible in every respect no mechanical work would be required in order to establish the conditions in the surface layers, since at every place in both columns (except between AB and CD) the pressure would be zero during the stretching, and K_n would always and everywhere be equal to P. Thus the work of separating the surface molecules from each other, as distinct from the work required to separate them from those of the other column, would be done entirely by P, i.e. by the heat supplied. In this case, therefore, the condition of strain in the surface layer itself would provide no mechanical energy to assist the spreading of a liquid. On the other hand, if the process of rupture is not completely reversible, the solid surface might itself possess available potential energy and therefore exert a tangential force on a spreading liquid. A possible mechanism by which this could be achieved is suggested later.

It is interesting to enquire how the above argument is modified if the ruptured column consists of an inextensible material (supposing that such could exist). In this case all the work is done *after* rupture, because a movement of the external forces before rupture would be contrary to the hypothesis of inextensibility. The work of separation will be completed when the width of the gap between the two new surfaces reaches the value c. The surface layers cannot be in a state of strain if the substance is inextensible, and the mutual repulsion of the molecules (P) cannot assist in the process of rupture because this is complete as soon as the distance between the molecules on either side of the gap exceeds its original value by only an infinitesimal amount. Superficial latent heat depends upon the existence of compressibility and therefore is zero in the case we are considering. If the substance is a liquid, surface tension exists, however, because an argument such as Bakker's is independent of whether or not the liquid is compressible. The work done in separating the liquid columns after rupture appears as surface energy. If the newly formed liquid surfaces are allowed to do work by contraction, the work derivable when the two columns of reduced cross-section are allowed to come together again will fall short of the original work of separation by the amount of work done by surface tension during the contraction of the surfaces. In other words, the attraction between the columns will have been decreased in proportion to the decrease of the area of the surfaces facing each other.

Once again the reasoning as to the existence of surface tension is dependent upon the fluid character of the substance and, therefore, does not apply to a solid. The only way in which the energy of the two separated columns of an inextensible solid can be recovered is by allowing them to come together again, and it seems hardly appropriate to associate such potential energy of separation with the *surfaces*. The newly formed surface layers may acquire stresses as a result of the removal of one column from the other but a stressed inextensible solid is incapable of performing work. The case of an inextensible solid is included here because, although it is hypothetical, it is shown in the next section that it is possible to create a very simple picture of the mechanism of spreading and capillary rise by treating the solid as inextensible.

§ 5. THE EQUILIBRIUM BETWEEN SOLID AND LIQUID SURFACES

In figure 9 the full lines represent the diagram which usually accompanies an investigation of the equilibrium between solid and liquid surfaces. By considering a displacement in the form of a spreading of the liquid over the solid surface and remembering that the condition for equilibrium is that a small displacement requires no external work, we write

$$\gamma_L \cos\theta = \gamma_S - \gamma_{SL}, \qquad \ldots\ldots(5)$$

where γ_L and γ_S are respectively the free surface energies of the liquid and solid against vapour and γ_{SL} is the free energy of the composite surface between solid and liquid. The dotted lines are inserted in figure 9 in order to indicate a less abstract state of affairs than that suggested by the unfinished diagram. We are concerned only with surface energies so that the effects of gravity can be ignored. The angle between the two solid surfaces is arranged so that the liquid/vapour surface is plane, and therefore the pressure in the liquid is equal to that in the vapour. The spreading can therefore be effected without the performance of work by injecting more liquid from outside through a tube which ends in the interior of the existing mass. This elucidation is, however, purely incidental.

The foregoing treatment is physically somewhat unsatisfying. The relation derived depends only upon the definition and conservation of free surface energy and therefore gives little information beyond the fact that there is, in general, an equilibrium value for the contact angle between a given solid and liquid. It is tempting to replace the surface energies by surface tensions and to regard the equation as an expression of a state of equilibrium between the three tensions acting at A (figure 9). Even cursory reflection at this stage, however, gives rise to doubts as to the justification for supposing that the solid/vapour surface assists spreading in the same way as would a liquid surface, i.e. by contracting and exerting a tension on A while doing so. Nevertheless the matter is usually left in this state.

In the following treatment it is assumed that the solid surface has no free energy, or more precisely, that its free energy is not modified when liquid replaces vapour in contact with the surface. We deal, therefore, with the inextensible solid discussed in the previous section and show that the existence of spreading and capillary rise can be explained on this basis. This is equivalent to saying that the occurrence of these phenomena does not in itself constitute any experimental evidence as to the existence of free energy in a solid surface.

Outside a solid surface there is a field of force directed perpendicularly inwards and extending beyond the surface for a distance equal to the maximum range of attraction of the solid molecules. When a liquid is in contact with such a surface this field will diminish or overcome the attraction towards the interior of the liquid which the surface molecules of the liquid experience and which normally gives rise to the tension in the liquid/vapour surface. The surface tension ($\gamma_{L(S)}$) of the liquid surface in contact with the solid is therefore less than that of the liquid/vapour surface and may, quite possibly, be negative, i.e. a surface pressure. Since we have assumed that the state of the solid surface is not modified by the presence of liquid in contact with it, it follows that the attractive force of the solid on a liquid particle at A (figure 9) must be normal to the surface and have no tangential component. Therefore the equilibrium of A is represented by the equation

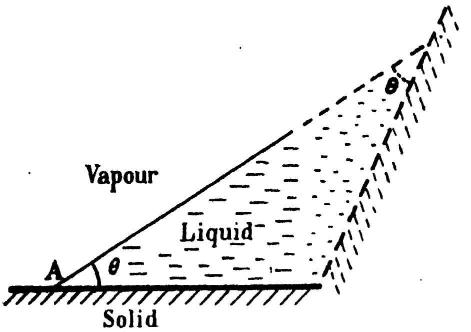


Figure 9.

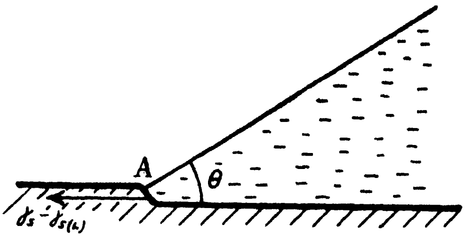


Figure 10.

$$\gamma_L \cos\theta + \gamma_{L(S)} = 0. \qquad \ldots\ldots(6)$$

Equation (6) shows that $\gamma_{L(S)}$ is positive when θ exceeds 90°, zero when θ is equal to 90°, and negative when θ is less than 90°. In the last case the liquid surface in contact with the solid is in a state of pressure tending to make the liquid spread.

If, as is quite possible, the solid and liquid are such that

$$\gamma_L < -\gamma_{L(S)},$$

equilibrium is not possible according to equation (6). The force opposing spreading ($\gamma_L \cos\theta$) is insufficient to balance the spreading force ($-\gamma_{L(S)}$) even when θ has become zero. The liquid therefore spreads until it forms a layer of molecular thickness in which, ultimately, there is no tension or pressure. The case of zero contact angle must not be considered as a special one such as would be represented by any given *finite* value of θ, but rather as being characteristic of the *class* of substances for which $\gamma_L \leqslant -\gamma_{L(S)}$.

The mechanism of spreading implied in this treatment is simply the creation of a surface pressure by the attraction of liquid molecules into the surface layer by the field of force of the solid. This is equivalent to the "squeezing-out" process envisaged by Leslie, the attraction of the solid upon liquid molecules beyond the first layer causing a tangential pressure in the surface layer. This idea is not popular among present-day surface specialists, but it must evidently be considered on the same footing as the simple theory which ascribes the free energy of a liquid/vapour surface to an inward attraction acting on the surface molecules. If an outward attraction is superimposed the result will be a reduced or negative surface energy.

If we no longer suppose that the free energy of the solid surface is unchanged when liquid spreads over it, then as a result of the replacement of vapour by liquid in contact with the solid surface, a certain amount of mechanical energy becomes available and the principle of virtual work leads to the equation

$$\gamma_L \cos\theta = \gamma_S - \gamma_{S(L)} - \gamma_{L(S)}, \qquad \ldots\ldots(7)$$

where $\gamma_{S(L)}$ is the free energy of unit area of the solid surface when covered with liquid. This is, of course, identical with equation (5), the energy of the composite surface (γ_{SL}) being expressed as the sum of the energies of the two surfaces of which it is composed. Since $\gamma_L \cos\theta$ and $\gamma_{L(S)}$ in equation (7) are real forces it follows that $\gamma_S - \gamma_{S(L)}$ must represent a real force acting tangentially on A, and it is necessary to seek a mechanism by which this force can be exerted on the liquid.

The solid can no longer be regarded as inextensible if the condition of its surface is modified by the nature of medium with which it is in contact. The extension in a direction normal to the surface, which, as already explained, is produced when the solid surface is formed by breaking a column, will be reduced by the presence of the liquid because the value of K_n in the surface layer of the solid will be increased, and for equilibrium this necessitates an increased value of P. There will, therefore, be a very small but definite "step" or change of level in the solid surface at A, as shown in figure 10, and the attraction which the solid exerts on a liquid particle at A is no longer normal to the general direction of the solid surface as it would be in the absence of the step. The tangential component of this attraction must, by equation (7), be equal to $\gamma_S - \gamma_{S(L)}$ and be directed away from the liquid.

It has previously been suggested in this lecture that the existence of free surface energy in a solid would indicate that the process of formation of the surface by the rupture of a column is irreversible. In these circumstances the wetting of the surface by a liquid would also be irreversible. This is a possible explanation of the well known observation that the advancing and receding contact angles differ considerably from each other.

§ 6. CAPILLARY RISE

The mechanism of capillary rise is, of course, closely related to that of spreading. It is simply spreading in a vertical direction which is eventually arrested by the force of gravity acting on the raised liquid.

A simple treatment of capillary rise can be presented by assuming, in the first place, that the walls of the tube or plates between which the liquid rises do not, themselves, exert a tangential force on the line of contact of the meniscus with them. That is to say, the surface energy of the solid is unaltered by wetting and can therefore be treated as inextensible. In this case the driving force is the surface pressure in the layer of liquid adjacent to the walls and equilibrium is established when the upward force due to this pressure is equal to the weight of the raised liquid column. Figure 11 is a representation of this state of affairs. The thick lines indicate surfaces which are in a state of tension and the dots show surfaces in a state of pressure. Inside the tube the surface pressure $-\gamma_{L(S)}$ supports the surface layer which covers the meniscus and this in turn supports the weight of the raised column by cohesion. Thus considering the equilibrium of the raised liquid and ignoring the weight of the liquid in the meniscus, we have

$$-2\pi a\gamma_{L(S)} = \pi a^2 hg\rho, \qquad \ldots\ldots(8)$$

where a is the radius of the tube, h is the height of the raised column and ρ is the density of the liquid. Furthermore, assuming (as is always done) that the magnitude of the angle of contact is independent of the action of gravity, we have, either by equation (6) or by considering the equilibrium of a liquid particle at the extreme edge of the meniscus,

$$-\gamma_{L(S)} = \gamma_L \cos\theta. \qquad \ldots\ldots(9)$$

Combining the last two equations gives the usual relation

$$\gamma_L \cos\theta = \frac{ahg\rho}{2}. \qquad \ldots\ldots(10)$$

If a very simple analogy may be permitted, the surface of the meniscus can be likened to a clothes line. The sum of the two vertical components of the tensions at the end of the line is equal to the weight of the clothes. This is, in fact, how the matter is very frequently treated in elementary text-books: the force $2\pi a\gamma_L \cos\theta$ acting upwards round the edge of the meniscus is equated to the weight of the column. This treatment is unsatisfying, however, unless it includes an explanation of how the edge of the surface layer of the meniscus (i.e. the ends of the clothes line) is supported. The foregoing argument shows that the surface pressure ($-\gamma_{L(S)}$) provides the support.

The case of zero contact angle must next be considered. It has already been explained that when $\gamma_L < -\gamma_{L(S)}$ the contact angle is zero and that for equilibrium a thin layer of liquid must cover all the exposed surface of the solid. Equation (9) is, of course, inoperative in these circumstances and so also is equation (8), because if it were applied as it stands, it would give different values of h for a given liquid with tubes of different materials ($\gamma_{L(S)}$ being dependent upon the nature of the solid), which is contrary to experience when θ is zero. It must be remembered, however, that the surface pressure acts upon the liquid column via the surface of the meniscus, so that, when the tension in this surface (γ_L) is less than $-\gamma_{L(S)}$, it is γ_L which decides the value of h (cf. the fact that the maximum weight which a clothes line can support is governed by its breaking tension). Thus the liquid initially rises to its equilibrium level given by

$$\gamma_L = \frac{ahg\rho}{2},$$

and after this the pressure in the layer adjacent to the walls causes the surface of the meniscus to extend indefinitely up the walls of the tube until, theoretically at any rate, the thin film of liquid so formed meets a similar film spreading up the outside wall of the tube from the liquid surrounding it. In the case we are considering, conditions are favourable to spreading, which means that the film of liquid on the walls of the tube above the meniscus is exerting a net force tending to extend its boundaries. Let this be ϕ per unit length of boundary. Then the equilibrium of a small element of liquid situated at the edge of the meniscus is given by

$$-\gamma_{L(S)}-\phi=\gamma_L. \qquad \ldots\ldots(11)$$

As to the value of ϕ, we can say that if the film is thick enough for its liquid/vapour surface to be beyond the influence of the attractive field of the solid, then the tension in this surface will be γ_L and the pressure in the liquid/solid surface of the film will be $-\gamma_{L(S)}$. Hence the net force which such a film exerts on unit length of its boundary will be the resultant of the pressure and tension in its two surfaces and will be given by

$$\phi=.-\gamma_{S(S)}-\gamma_L, \qquad \ldots\ldots(12)$$

which is identical with equation (11). It is not contended that the film is necessarily thick enough to possess two independent surfaces—merely that the condition for equilibrium (equation (11)) requires that the value of ϕ shall be the same as it would be if the film were of this nature.

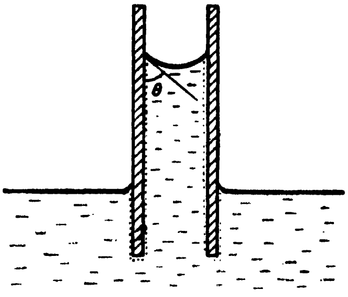


Figure 11.

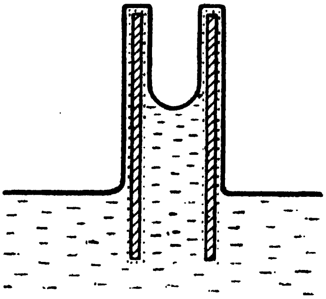


Figure 12.

Thus we can visualize the equilibrium of the system in the way indicated by figure 12, where, as before, the heavy lines represent surfaces in a state of tension γ_L and the dots represent the surface layer which is exerting a pressure $-\gamma_{L(L)}$. It will be seen from the diagram that, although the pressure $-\gamma_{L(S)}$ provides the initial raising force, it is not concerned in the final equilibrium which can be regarded as existing between the tension γ_L and the weight of the raised column. The film covering the walls of the tube above the meniscus is exactly equivalent, physically, to a soap film contained in a vertical frame dipping into a dish of liquid.

In practice, when a capillary tube is placed in a liquid the state of affairs depicted in figure 12 is, no doubt, never realized, because even when θ is zero the spontaneous spreading of a liquid over a solid surface does not, as a rule, take place. Nevertheless there will be some creeping of the liquid up the walls of the tube above the column—indeed, the method of ensuring zero contact angle is to wet the inside of the tube above the ultimate equilibrium position of the meniscus. The film thus formed may terminate at some place on the walls where spreading is arrested by an irregularity or variation in the nature of the solid surface but, for equilibrium, the film must nevertheless exert a downward force on the edge of the meniscus equal to $-\gamma_{L(S)}-\gamma_L$ per unit length. In many of the standard capillary-rise determinations, the tube and reservoir are contained in a closed vessel. In these circumstances if the walls of the tube were initailly wetted the likelihood of a stable film persisting indefinitely is much greater than when the tube is exposed to the air. The condition of the film would be influenced by two factors in addition to those already mentioned. In the higher parts of the tube it must be in equilibrium with a slightly lower vapour pressure, and also the net pressure ϕ which it exerts would be less on account of the variation of hydrostatic pressure in a vertical direction within the film itself.

It is evident that even if some mechanical energy is made available when the solid surface is wetted, this fact does not affect the above discussion of the equilibrium when the contact angle is zero. If the angle is not zero, however, we do not envisage a liquid film covering the walls of the tube above the meniscus, and any difference which may exist between the surface energies of the dry and wet solid surfaces must be written down as an upward force $\gamma_S-\gamma_{S(L)}$ acting on unit length of the edge of the meniscus in addition to $-\gamma_{L(S)}$. The relation between γ_L and h is, of course, unaltered by this modification.

§ 7. ACKNOWLEDGMENTS

The author wishes to thank Professor E. N. da C. Andrade, F.R.S., and Mr. D. O. Wood for reading the first draft of this lecture and for making valuable suggestions with regard to the method of presentation.

DISCUSSION

Dr. L. HARTSHORN. I am surprised to hear that so many reputable physicists have had the temerity to assert that surface tension is fiction and surface energy real. Personally I doubt whether the word " real " has any place in physics, which I take to be the science of measurable properties. If by " real " you mean measurable, and by "fiction " not measurable, then you must admit that the surface tension of liquids is real because it is directly measurable, while surface energy is much more doubtful. Energy is of such dominating importance in unifying the various branches of physics, that we are apt to forget that it is not directly measurable, and to assume that its high status in theory necessarily gives it the same status in practice. A striking example occurs in electricity. Any elementary student, given a good potentiometer, can measure voltage with ease to an accuracy of say 0·01 per cent or better. This quantity is usually stated to be defined in terms of the energy expended in moving an electric charge from one place to another, a quantity which even the most accomplished experimentalist could not measure with an accuracy of 1 per cent. Is this good physics ? Which is the real quantity ? As we are discussing surface tension and not electricity I will only mention that Dr. Norman Campbell and I struck this oddity when making a critical examination of electricity, which we hope will appear in the *Proceedings*. Returning to surface tension, I hold that those who wish to make such distinctions may well say that in liquids surface tension is more real than surface energy, because it is more directly measurable : in solids, however, the tension is no longer directly measurable ; the two conceptions are on much the same footing, and surface energy may be preferred because energy is a more powerful conception in theory. I welcome the spirit of Dr. Brown's lecture in refusing to accept with the customary complacency traditional notions that are not satisfying: the interest it has aroused is a most encouraging sign.

Mr. C. GURNEY. I could not more agree with Dr. Brown in his criticism of the idea that the tension in surface of a liquid is a mathematical fiction. I believe it is as real as the tension in a piece of stretched wire. I would like to suggest, however, a rather different explanation of this tension.

The environment of the atoms in the surface of a liquid is different from that of those in the interior. The surface atoms have fewer nearest neighbours and consequently they are less strongly held to one another. It is therefore easier for thermal motions to eject atoms from the surface than from the interior; in other words, the tendency for atoms to escape from the surface is greater than the tendency for atoms to escape from the interior. This tendency of atoms to escape is measured by the chemical potential. The chemical potential of surface atoms is therefore higher than that of interior atoms. In equilibrium the chemical potential of the surface atoms must equal that of the interior atoms, and of course also that of the atoms in the vapour phase. How can this be achieved? It may be convenient here to digress and consider other cases of thermal equilibrium. Consider a pure substance, say lead. At atmospheric pressure the melting-point of lead is above room temperature. The chemical potential of solid lead is therefore less than that of liquid lead, so that the solid phase is stable. How could we have solid and liquid lead in equilibrium at room temperature. The answer is, that a suitable stress system must be applied, either to the solid or liquid or both. If we apply a hydrostatic pressure to the solid we could increase the chemical potential of the solid until it equalled that of the stress-free liquid. Two-phase equilibrium would then be possible. Or we could apply a hydrostatic tension to the liquid. This would decrease the chemical potential of the liquid until it equalled that of the stress-free solid. This gives the clue to the question of surface-liquid equilibrium. We can reduce the chemical potential of the surface by applying a tension to it. Tension reduces the tendency of atoms to escape from the surface, and at any given temperature there will be one value of the tension which reduces the tendency of atoms to escape from the surface so that it is equal to the tendency of atoms to escape from the liquid. Liquid and surface then have the same chemical potential and equilibrium is attained. For a liquid under hydrostatic pressure $\partial\mu/\partial p = \bar{V}$, where μ is chemical potential, p is pressure, V is specific volume. For a surface under tension $\partial\mu/\partial\gamma = -\bar{A}$, where γ is tension in surface and $\bar{A}$ is specific area.

A freshly formed surface having a given perimeter might have no tension in it. Atoms therefore escape from the surface faster than they enter it, until the spacing of atoms in the surface is further apart than the stress-free spacing. A tension is thus set up in the surface. It is not a dynamic tension as is thought to occur in rubber. It is just the same sort of tension as in a stretched wire, only unlike the tension in a stretched wire, the tension in the surface can never relax by creep. As soon as it tended to do so, more atoms would escape from the surface than would enter it, and the surface would be tightened up. There is no need to assume that surface atoms have any special rigidity or orientation. The liquid must have a surface and the surface cannot be stable unless it has a tension acting in its plane. Here, in a mobile liquid, is a tension that cannot relax. In an unstrained liquid the surface molecules leave the surface and enter the interior until a tension can be set up and equilibrated by a pressure within the liquid. The surface thus becomes spherical. The fact that the liquid in a drop is under pressure, causes the chemical potential of the liquid to be higher than that under a flat surface. A smaller tension in the surface is then sufficient to reduce the chemical potential of the surface to equal that of the liquid. Surface tension thus decreases as the radius becomes smaller. By equating the partial derivatives of the chemical potential of the liquid with respect to pressure to that of the surface with respect to tension we obtain a formula for the change of surface tension with radius. It is $\partial\gamma/\partial r = (2\gamma/r^2)(\bar{V}/\bar{A})$, where γ = surface tension, r = radius, $\bar{V}$ = specific volume of liquid and $\bar{A}$ = specific area of surface. $\bar{V}/\bar{A}$ is thus the effective thickness of the surface film. This equation does not seem to be generally known, but it is really a particular case of a more general equation given by Gibbs in 1876 (*Collected Works*, Longmans, New York, 1928, p. 232).

I would now like to refer to the surface tension of solids. If the solid is slowly cooled from the liquid state without phase change—for example, freshly made glass—it might have a tension in the surface approximating to the equilibrium value. Owing to viscosity, it would probably never attain the true equilibrium value, but it would tend to it for very slow cooling. If, on the other hand, the surface is formed by fracture, it would have no tension (except the polarization tension) in the surface to begin with, and would achieve its equilibrium tension so slowly that appreciable tension might never be developed. In this case the chemical potential of the surface layer would be higher than that of the underlying material, and the surface layer would be more chemically reactive.

In conclusion I would like to mention the tension in the surfaces of solids caused by polarization of ions and analogous effects. Surface atoms have fewer nearest neighbours, and consequently the bond strength between a surface atom and each of its remaining neighbours is greater than that between an interior atom and each of its greater number of neighbours. For example, in a covalent bonded material such as diamond, the bonds which are broken when the surface is formed, become resonating double bonds between the surface atoms and their remaining neighbours. The strength of a carbon double bond is greater than that of a single bond and the equilibrium atomic spacing is smaller. The surface atoms would therefore tend to move closer together, but they are prevented from doing so by the rigidity of the underlying solid. A tension is therefore set up in the surface. In an ionic bonded material such as sodium chloride, the surface ions become polarized, the concentration of electric charges becoming increased in the directions of the neighbouring ions and again the bond strength is increased. It is important to realize that the polarization tension is not the equilibrium surface tension, although it will reduce the chemical potential of the surface atoms and will thus reduce the additional tension which must be applied to achieve equilibrium.

AUTHOR'S REPLY. It is pleasing to hear that both Dr. Hartshorn and Mr. Gurney agree with me in according a physical reality to surface tension. I am inclined to believe that those who profess to regard it as a fiction are not always really convinced by their own contention.

In his opening remarks Dr. Hartshorn implies that " reality " is a rather ill-defined concept and I entirely agree. In this lecture I have not been concerned with the philosophical question of the relative reality of force and energy in general. My sole purpose has been, so to speak, to ask those who concede a physical existence to surface energy, but not surface tension, whether they have any more reason for doing this than for regarding, say, the weight of a body as a fiction, but its gravitational potential energy as a reality. In other words, I believe that surface tension and surface energy are no more or less real than any other types of force and energy.

Dr. Hartshorn seems inclined to estimate the reality of physical quantities by the directness with which their values can be determined in particular cases. If the application of this criterion would make velocity less real than the length and time from which it is derived (except when it is read directly from the speedometer of a car !) I should be inclined to question its suitability.

Personally I doubt whether much is to be gained by trying to ascribe degrees of reality to force and energy in liquid and solid surfaces as Dr. Hartshorn suggests later on in his remarks. It seems to me that force is force and energy is energy wherever they are found and whatever may be their physical " cause ".

Mr. Gurney's remarks present the mechanism of surface tension in an interesting light. In his equation for the variation of surface tension with drop radius I wonder whether it is not necessary also to take into account the variation of vapour pressure with curvature. The possibility mentioned by Mr. Gurney of a difference of condition between solid surfaces, formed on the one hand by fusion and subsequent solidification and on the other by fracture, suggests that measurements of contact angle between a given liquid and the two types of surface might yield interesting results.

ULTRA-VIOLET ABSORPTION BAND-SYSTEMS OF PbO, PbS, PbSe AND PbTe

By E. E. VAGO and R. F. BARROW,
Physical Chemistry Laboratory, University of Oxford

MS. received 22 *October* 1946

ABSTRACT. A study has been made of the absorption spectra of PbO, PbS, PbSe and PbTe in the ultra-violet region ($\lambda > 2050$ A.). The vibrational analysis of new band-systems of PbS, PbSe and PbTe are reported, together with some observations on systems of PbO ($\nu_e \sim 35000$) and of PbSe ($\nu_e \sim 24818$ cm.$^{-1}$) already known. The table below includes the values of ν_e, ω_e and $x_e\omega_e$ at present available for the upper states of these molecules which are involved in absorption systems in this region of the spectrum.

State	PbO	PbS	PbSe	PbTe
G	—	—	—	46541·7 159·6 1·4
F	—	47770 370 (7·8)	45220·9 224·8 0·50	41658·8 176·4 1·0
E	34900 430 —	? 34000 — —	—	—
D	30198·7 530·5 2·92	29650·5 299·3 1·57	28418·0 190·4 0·53*	27176·5 142·6 1·58

* $y_e\omega_e = -0.004$.

§ 1. INTRODUCTION

The electronic states of the group of molecules comprising the oxides, sulphides, selenides and tellurides of C, Si, Ge, Sn and Pb were briefly reviewed a little while ago (Barrow, 1944): one conclusion was that several hitherto undetected states should exist, some of which should be involved in absorption band-systems lying in the near infra-red and in the ultra-violet regions of the spectrum. The ultra-violet region has since proved to be surprisingly fruitful, and we have recently described systems of the silicon and tin compounds occurring there (Vago and Barrow, 1946 a, b). In the present paper we give the results of a study of the absorption spectra of the lead compounds in the same region.

The arbitrary low-wave-length limit of about 2050 A. set by our present experimental arrangements implies that the work is far from complete. For example, we have found levels of PbTe at 41660 (F) and 46540 (G) cm.$^{-1}$: since, for corresponding systems, ν_e very often shifts to higher values with decreasing atomic number, we are led to predict an F level of PbO and G levels of PbO, PbS and PbSe which should take part in absorption systems in the region 50 000 to 65 000 cm.$^{-1}$ There is the same kind of assurance that SiO, GeO, SnO, SnS and SnSe will also absorb here, and somewhat less certain predictions may be made about many of the other molecules of this group. It is hardly possible to give an adequate discussion of the upper states of these molecules when so much experimental information is still lacking: for this reason we shall be concerned here primarily with the observations relating to the region 2050 to 3800 A.

§ 2. RESULTS

PbO

The absorption spectrum of PbO was studied in detail by Howell (1936), who recognized five band-systems and suggested the existence of a sixth. Of these, the system of shortest wave-length (called *F*←*X* by Howell, here E←X; see note to table 9) contained only seven bands. We therefore thought it of interest to examine this system again in the hope of extending it: at the same time we looked for absorption at somewhat shorter wave-lengths. We found no new bands in the range 2400–2700 A., but did obtain spectrograms of the E←X system which have led us to a rather different interpretation of this system.

The spectra were photographed in a large quartz Littrow instrument (dispersion about 3·3 A./mm. at 2800 A.) on Ilford Process plates. The effective path-length was estimated at 15 cm. and the temperature at about 1300° C. Bands of the D–X system of PbO are well developed under these conditions (see reproduction, figure 1) and at the short-λ end it is overlapped by the E←X system. Rotational structure lines, presumably of bands of the E←X system, can be traced to about 2550 A., but the strongest heads lie in the range 2800 to 3075 A.

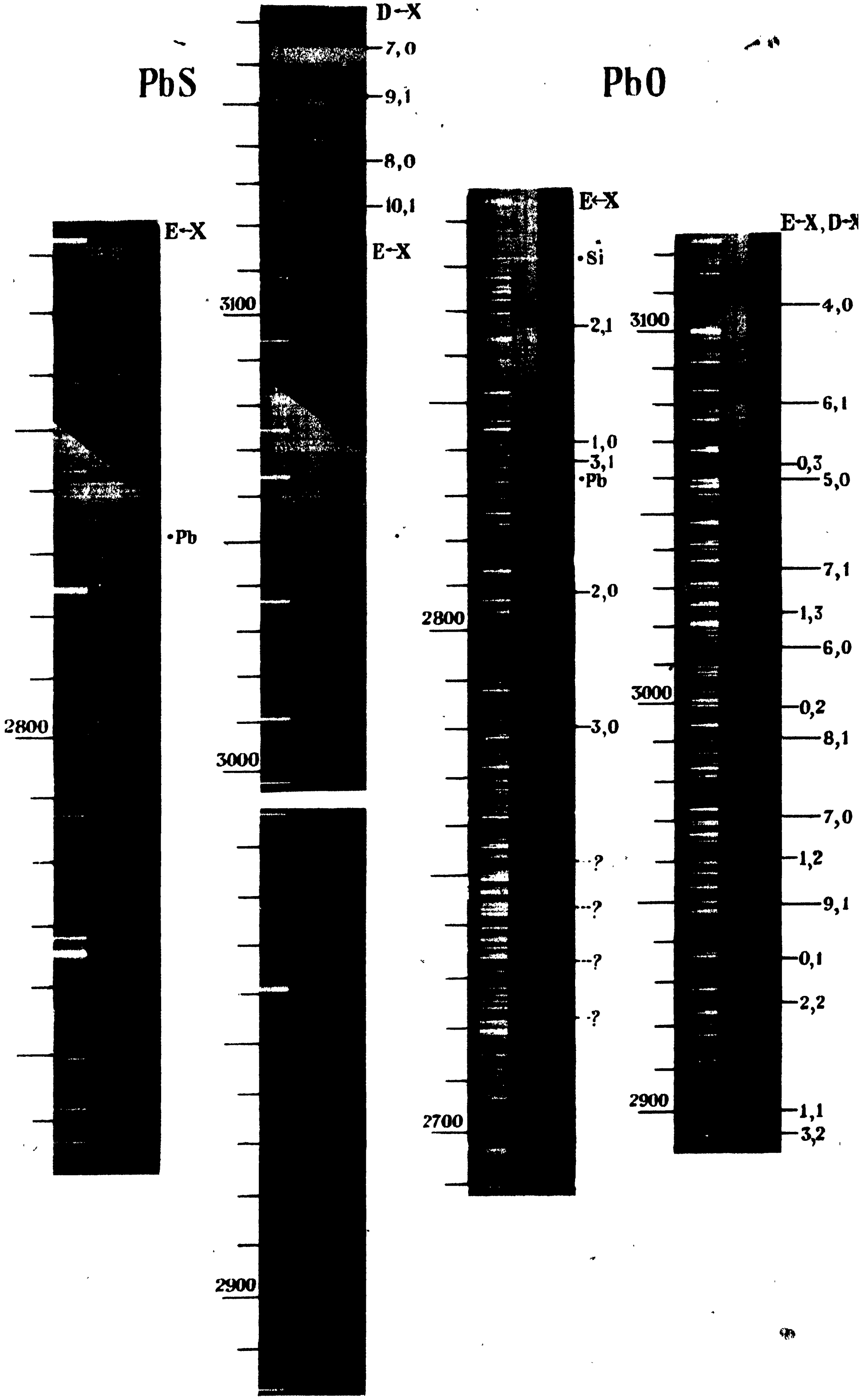
Our band-head measurements are summarized in the Deslandres scheme in table 1. The arrangement there differs from that given by Howell by the inclusion

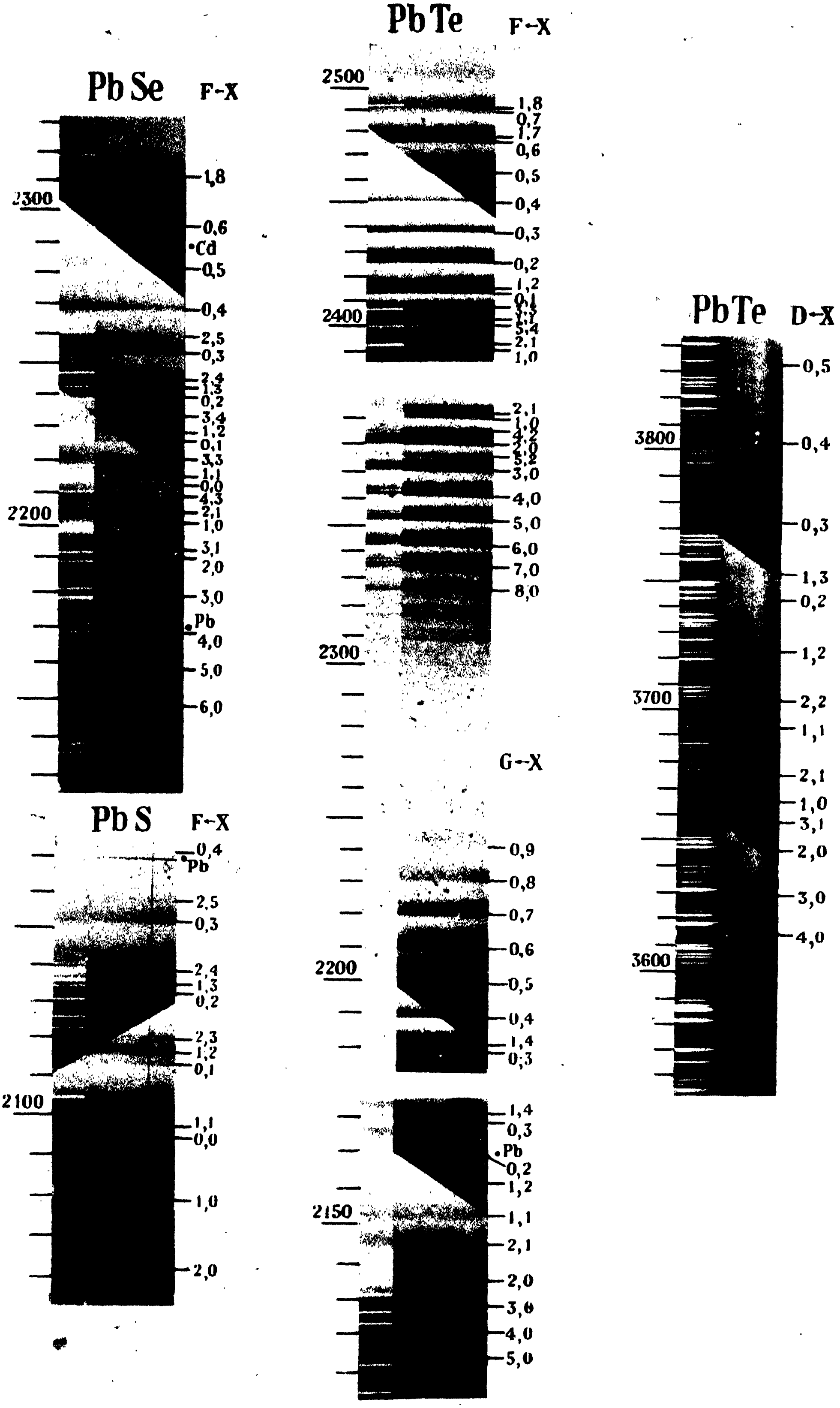
Table 1. Measurements of band heads of of the E←X system of PbO

Figures in parentheses are visual relative intensities.

3	(1)* 35959 *2780·1$_8$*	715·$_8$	(1)* 35243·5 *2836·57*	701·9	(1)* 34541·6 *2894·21*		
	363		364·0		371·0		
2	(1)* 35596 *2808·5*	716·$_8$	(5)* 34879·5 *2866·17*	708·9	(3)* 34170·6 *2925·64*		
	414		409·4		405·0		
1	(0) 35182 *2841·5*	712	(4) 34470·1 *2900·21*	704·5	(3) 33765·6 *2960·73*	699·6	(2)* 33066·0 *3023·38*
			422·3		425·5		424·2
0 ↑ v'			(2) 34047·8 *2936·19*	707·7	(4) 33340·1 *2998·52*	698·3	(4) 32641·8 *3062·67*
v''→0			1		2		3
$\Delta G''$ (mean):		714·7		705·8		699·0	
$\Delta G''$ (calc.):		714·4		707·0		699·6	

* Band also measured by Howell (1936).


PbS
E←X
•Pb
2800
D←X
7,0
9,1
8,0
10,1
E←X
3100
3000
2900
PbO
E←X
•Si
2,1
1,0
3,1
•Pb
2,0
2800
3,0
?
?
?
?
2700
E←X, D←X
4,0
3100
6,1
0,3
5,0
7,1
1,3
6,0
3000
0,2
8,1
7,0
1,2
9,1
0,1
2,2
2900
1,1
3,2


PbSe F←X
2300
2200
1,8
0,6
•Cd
0,5
0,4
2,5
0,3
2,4
1,3
0,2
3,4
1,2
0,1
3,3
1,1
0,0
4,3
2,1
1,0
3,1
2,0
3,0
•Pb
4,0
5,0
6,0
PbTe F←X
2500
2400
2300
1,8
0,7
1,7
0,6
0,5
0,4
0,3
0,2
1,2
0,1
3,3
1,1
5,4
2,1
1,0
2,1
1,0
4,2
3,2
2,0
5,2
3,0
4,0
5,0
6,0
7,0
8,0
G←X
2200
2150
0,9
0,8
0,7
0,6
0,5
0,4
1,4
0,3
1,4
0,3
•Pb
0,2
1,2
1,1
2,1
2,0
3,0
4,0
5,0
PbS F←X
2100
0,4
•Pb
2,5
0,3
2,4
1,3
0,2
2,3
1,2
0,1
1,1
0,0
1,0
2,0
PbTe D←X
3800
3700
3600
0,5
0,4
0,3
1,3
0,2
1,2
2,2
1,1
2,1
1,0
3,1
2,0
3,0
4,0

of the bands with $v'=0$ and 1 (the 1,3 band was also measured by Howell but listed as unclassified). The two sets of measurements on the bands with $v'=2$ and 3 are reasonably concordant, but the intensities differ qualitatively. Thus for the two bands at 2780 and 2866 we find (roughly) 1 and 5, while Howell quotes 8 and 8: again we find very little evidence for a head at 2753 listed by Howell as occurring with intensity 6. These discrepancies suggested that some of the bands might be due to impurities, for it seems difficult to account for such variations in intensity on the basis of any plausible differences in temperature or in plate-sensitivity. We have not, however, been able to attribute any of the bands to another molecule.

It seems probable therefore that the new bands with $v'=0$ and 1 do belong to the E←X system. This view, and the correctness of the analysis given in table 1, are to some extent confirmed by the $\Delta G''_v$ values derived therefrom, which are seen to be very close to those calculated from Howell's expression determined from measurements of the other, more extensive, systems of PbO. The $\Delta G'_v$ values decrease irregularly and unusually rapidly as v' increases, suggesting strong perturbation of the E level.

PbS

The absorption spectrum of PbS has been investigated by Rochester and Howell (1935): their measurements extend to about 3135 A.—the short-λ end of the system with $\nu_e = 29650$ (D–X, see note to table 9). We have now examined, and found bands at, both extremes of the region 2060–3200 A. Ilford Q1 plates and a Hilger medium quartz prism instrument (E.498: dispersion about 5·5 A./mm. at 2150 A.) were used for the range 2060–2200 A.: the range 2500–3100 A. was photographed in a first order of a 2·4-m. concave grating (linear dispersion about 7·4 A./mm.) and in the large quartz Littrow spectrograph on Ilford Process plates.

We have not been able to analyse the bands lying between 2750 and 3100 A. (table 2). The fact that many of them can be fitted into short progressions which yield differences (415–430 cm.$^{-1}$) of the order of the $\Delta G''_v$ values for PbS strongly suggests that they are indeed to be attributed to this molecule, but no likely

Table 2. Measurements of band-heads observed in the absorption spectrum of PbS

λ, I.A.	ν, cm.$^{-1}$	λ	ν	λ	ν
2754·85	36389	2852·0	35053	2974·35	33611*
57·8	350	63·7	34910*	76·8	583*
64·05	168	69·6	838*	86·8	471*
70·1	089*	70·9	822	97·15	355*
77·4	35994*	81·6	693*	3013·1	179*
84·65	901	92·25	565	24·15	057*
86·15	881	2904·6	418*	41·1	32873
93·95	781	15·7	287*	52·0	756*
2802·65	670*	30·6	113	62·8	640*
10·3	573*	41·15	33990*	71·8	545
21·25	435	52·15	863*	83·0	427
30·15	323*	59·25	782*	90·25	350*
41·55	182	64·65	721	3191·05	238*

* Indicates that the band may be fitted to a short v''-progression.

vibrational scheme has suggested itself; nor does it seem probable that all the bands can arise from high-v' transitions in the D–X system (which appears strongly in the plates). We are inclined to believe that some at least of the bands form an E←X system of PbS similar to that of PbO just described. At the short-λ end the bands crowd very closely together and have not been observed below about 2715 A., which may give a low-energy limit (~4·55 ev.) for a dissociation process.

The measurements and vibrational analysis of the bands in the region 2060–2200 Å. (see figure 1) are given in table 3. The system (F←X) has not been observed in its entirety, as we were unable to photograph below about 2060 A., but enough has been measured—some twenty bands—to make the vibrational analysis reasonably certain. The v''-assignments follow from the known values of $\Delta G''_v$ (taken from Rochester and Howell), and there is then little doubt about the values of v' as the system is bounded by an intense progression with $v' = 0$. The upper-state vibrational levels appear to be rather irregular: the expression given in table 9 has been derived on the assumption that the heads of bands with $v' = 0$ are displaced so as to lie about 12 cm^{-1} too low. This expression fits the values for the other band-heads satisfactorily, but it may be criticized on the ground that the value of $x_e\omega_e$ is abnormally large (7·8). It is possible therefore that the perturbations extend to some of the other vibrational levels of the F state: this question could probably be elucidated by extending the observations to shorter wave-lengths.

Table 3. Measurements of band-heads of the F←X system of PbS

4					48215 *2073·4*	+422	47792·7 *2091·70*						
					312		311						
3	§		48325 *2068·6*	+422	47903 *2086·9*	+421	47482 *2105·4*						46228 *2162·5*
			320				323						
2	48430 *2064·2*	+425	48005 *2086·8*		—		47158·8 *2119·82*	+412·3	46746·5 *2138·53*	+418	46329 *2157·8*		
	338		338				331·5		341				
1	48092 *2078·7*	+426	47666·5 *2097·24*	+421·4	47245·1 *2115·95*	+417·8	46827·3 *2134·83*	+421	46406 *2154·2*				
	363		362·4		366·9		371·2		367				
0 ↑ v'	47729·5 *2094·47*	+425·4	47304·1 *2113·31*	+425·9	46878·2 *2132·51*	+422·1	46456·1 *2151·89*	+417	46039 *2171·4*				
$v'' \rightarrow$	0		1		2		3		4		5		6

§ The limit of the observations was about 2060 A.

PbSe

One ultra-violet system of PbSe (D–X) has been described (Barrow and Vago, 1944). This was first obtained in emission, but spectrograms have since been taken in absorption, using the Hilger medium quartz instrument. Measurements of the new plates have confirmed the original vibrational analysis and have enabled several additional bands to be assigned (table 4). The expression for G'_v has been re-calculated and the result is given in table 9. Bands can be traced on the plates down to about 2870 A., which gives a minimum value for a dissociation limit at 4·3 ev. above the ground state.

Absorption bands have also been found in the region 2150–2300 A. (figure 2). These form a single system, F←X, whose vibrational analysis is given in table 5.

Table 4. Additional or amended measurements of bands of the D—X system of PbSe

v', v''	ν, cm^{-1}	v', v''	ν, cm^{-1}
14,1	30637	8,1	29585·5*
13,1	468	6,1	214·4*
14,2	364	4,0	121·6*
12,1	297	2,0	28751·9*
13,2	193	3,1	658·8*
11,1	120	0,0	373·8*
12,2	016	1,1	287·2*
10,1	29944	2,2	202·5*
11,2	845	0,1	098·2*
12,3	749	1,2	011·6*

* Band also observed in emission.

Table 5. Measurements of band-heads of the F←X system of PbSe

5	46309 *2158·7$_5$*				45747 *2185·2$_5$*										
	219														
4	46090 *2169·0*	277	45813 *2182·1*				45257 *2208·9*				44711 *2235·9*				
	220		221				224								
3	45870 *2179·4*	278	45592 *2192·7*				45033 *2219·9*	274	44759 *2233·5*						
	223		225						220						
2	45647 *2190·0$_5$*	280	45366·9 *2203·56*	278·0	45088·9 *2217·15*				44539 *2244·5*	270	44269 *2258·2*				
	228		225·1		221·8										
1	45419·3 *2201·02*	277·5	45141·8 *2214·55*	274·7	44867·1 *2228·11*	274·8	44592·3 *2241·84*								43503 *2298·0*
	224·3		223·7		224·8		223·4								
0 ↑ v'	45195·0 *2211·94*	276·9	44918·1 *2225·58*	275·8	44642·3 *2239·33*	273·4	44368·9 *2253·13*	273·6	44095·3 *2267·11*	273·6	43821·7 *2281·27*	269·0	43552·7 *2295 36*		
v''→	0		1		2		3		4		5		6		7

PbTe

Several unsuccessful attempts were made to excite a D–X system in emission, using methods which had proved effective for SnSe and PbSe. We then examined the absorption spectrum and found not only the expected D←X system, but two other systems, F←X and G←X, as well.

The D←X system lies in the region 3550–3900 A. (figure 2): it was photographed in a first order of the grating instrument. The vibrational analysis is straightforward and is given in table 6.

Table 6. Measurements of band-heads of the D←X system of PbTe

6	27933 *3579·0*													
	125													
5	27808 *3595·1*													
	129													
4	27679 *3611·8*													
	131													
3	27548·0 *3628·99*	210·9	27337·1 *3657·00*											
	130·2		131·1											
2	27417·8 *3646·23*	211·8	27206·0 *3674·62*	211·7	26994·3 *3703·44*	209·0	26785·3 *3732·33*							
	137·4		137·4		137·5		140·1							
1	27280·4 *3664·60*	211·8	27068·6 *3693·27*	211·8	26856·8 *3722·39*	211·6	26645·2 *3751·96*	212·0	26433·2 *3782·05*					
	137·7		138·2		138·0		135·7		134·3					
0 ↑ v'	27142·7 *3683·19*	212·3	26930·4 *3712·23*	211·6	26718·8 *3741·62*	209·3	26509·5 *3771·16*	210·6	26298·9 *3801·37*	212·6	26086·3 *3832·35*	212·2	25874·1 *3863·78*
	$v''\rightarrow$ 0		1		2		3		4		5		6	

The F←X and G←X systems were photographed in the medium quartz prism spectrograph. The bands of the F←X system (figure 2) lie between 2300 and 2500 A.: again there is little doubt about the vibrational analysis (table 7), but the head of the 6,0 band appears to be displaced about 12 cm⁻¹ from its expected position.

The strongest bands of the G←X system (figure 2, table 8) form a $v' = 0$ progression in the region 2160–2250 A. Altogether seventeen bands have been assigned to this system: it is likely that more could be observed at shorter wavelengths, which would make possible the derivation of a more accurate expression for G'_v than that given in table 9, but it does not seem probable that the analysis would be changed thereby.

§ 3. SUMMARY

The results are summarized in table 9. It is not proposed to discuss them in detail, but it is perhaps of interest to compare them with those for the corresponding tin compounds. This is done in table 10, which includes the values of ν_e and of the force-constants expressed as percentages of those in the ground states. There seems to be some evidence that the latter values are grouped significantly, i.e. that the states labelled with similar letters are indeed spectroscopically similar, but perhaps the most useful function of this kind of table at the moment is to demonstrate the gaps in our knowledge. However, all the spaces in table 10 do not necessarily imply the existence of unknown stable states: thus, it may be that B←X systems of PbSe and PbTe do not exist, for the dissociation energies of the E states of the tin compounds are probably rather small, decreasing from oxide to telluride (Vago and Barrow, 1946 b), so that the corresponding levels of PbSe and PbTe may be unstable. This argument does not apply to the other states in

Table 7. Measurements of band-heads of the F←X system of PbTe

$v' \uparrow$ / $v'' \to$	0		1		2		3		4		5		6	
12					43181 *2315·1*									
					156									
11			43236 *2312·1$_5$*	211	43025 *2323·5*									
			155											
10	43293 *2309·1$_5$*	212	43081 *2320·5*											
	159		160											
9	43134 *2317·6$_5$*	213	42921 *2329·1$_5$*											
	159		158											
8	42975 *2326·2*	212	42763 *2337·7$_5$*											
	156													
7	42819 *2334·7*													
	150													
6	42669·4 *2342·88*													
	172·8													
5	42496·6 *2352·41*				42067·2 *2376·42*				41643·9 *2400·58*					
	168·1				166·9									
4	42328·5 *2361·75*				41900·3 *2385·89*	211·1	41689·2 *2397·97*				41270·0 *2422·33*			
	169·9						168·8							
3	42158·6 *2371·27*	212·1	41946·5 *2383·26*				41520·4 *2407·72*						40889·6 *2444·87*	
	169·4		169·6											
2	41989·2 *2380·84*	212·3	41776·9 *2392·94*						41142·1 *2429·86*	210·0	40932·1 *2442·33*			
	172·0		172·1								165·6			
1	41817·2 *2390·63*	212·4	41604·8 *2402·83*	212·2	41392·6 *2415·15*	209·3	41183·3 *2427·43*				40766·5 *2452·25*	210·5	40556·0 *2464·98*	211·1
			177·5		175·7		177·5				179·9		177·5	
0			41427·3 *2413·13*	210·4	41216·9 *2425·45*	211·1	41005·8 *2437·94*	209·5	40796·3 *2450·46*	209·7	40586·6 *2463·12*	208·1	40378·5 *2475·82*	207·7

$v' \uparrow$ / $v'' \to$	7		8
1	40344·9 *2477·88*	211·5	40133·4 *2490·94*
	174·1		
0	40170·8 *2488·62*		

the table, however, and if the arrangement is correct, nine more F←X and G←X systems of these molecules await discovery.

Table 8. Measurements of band-heads of the G←X system of PbTe

4	47126 *2121·3*													
	151													
3	46975 *2128·1*	213	46762 *2137·8*											
	154		151											
2	46821· *2135·1*	210	46611 $2144{\cdot}7_5$											
			155											
1	—		46456 *2151·9*	207	46248·7 *2161·54*	208·5	46040·2 *2171·33*	208	45832 *2181·2*					
					158		159·5		161					
0 ↑ v'	—		—		46091 $2168{\cdot}9_5$	210	45880·7 *2178·88*	209·8	45670·9 *2188·89*	212·2	45458·7 *2199·11*	208·8	45249·9 *2209·26*	209
$v'' \rightarrow$	0		1		2		3		4		5		6	

0 ↑ v'	45041 *2219·5*	210	44831 *2229·9*	212	44619 *2240·5*
$v'' \rightarrow$	7		8		9

Table 9. Values of ν_e, ω_e and $x_e\omega_e$

State	PbO	PbS	PbSe	PbTe
G	—	—	—	46541·7§ 159·6 1·4
F	—	47770 370 (7·8)	45220·9§ 224·8 0·50	41658·8§ 176·4 1·0
E	34900‡ 430 —	?34000 — —	—	—
D	30198·7 530· 5 2·92	29650·5‡ 299·3 1·57	28418·0 190·4 0·53*	27176·5 142·6 1·58
X	0 721·8 3·70	0 428·1 1·20	0 277·4 0·56	0 211·8 0·12

* $y_e\omega_e = -0{\cdot}004$.

‡ It has seemed best to give the same letters to states which appear to be analogous. Unfortunately, this has led to two alterations. For PbO, the level with $\nu_e \sim 35000$ was formerly F: it is doubtful, however, if the level formerly known as E really exists (Howell, 1936), and we prefer this letter for the 35000 state as it emphasizes some similarities with the E levels of SnO, etc. The level of PbS at 29650 (formerly called E) is probably analogous to what are already known as the D levels of PbO and of PbSe, and it is has, therefore, been called D above.

§ A brief notice of the results of vibrational analysis of these systems has recently been published by Sharma (1946). His constants for the F←X systems of PbSe and PbTe agree quite closely with ours, but he obtains $\nu_e = 45918$ for the G←X system of PbTe.

Table 10. Values of ν_e (in ev.) and of the force-constants expressed as percentages of those in the ground states

State	SnO	PbO	SnS	PbS	SnSe	PbSe	SnTe	PbTe
G	—	—	—	—	—	—	5·86 78	5·77 56·8
F	—	—	—	5·92 75	5·93 77	5·61 65·7	5·46 78·4	5·16 69·4
E	4·56 36·6	4·33 35·5	4·10 36·4	? 4·2 —	3·81 35·1	— —	$3{\cdot}4_7$? 33	— —
D	3·68 50·2	3·74 54·0	3·51 46·3	3·68 48·9	3·42 46·2	3·52 47·1	3·15 47·6	3·37 45·3

REFERENCES

Barrow, R. F., 1944. *Proc. Phys. Soc.*, **56,** 204.
Barrow, R. F. and Vago, E. E., 1944. *Proc. Phys. Soc.*, **56,** 78.
Howell, H. G., 1936. *Proc. Roy. Soc.*, A, **153,** 683.
Rochester, G. D. and Howell, H. G., 1935. *Proc. Roy. Soc.*, A, **148,** 157.
Sharma, D., 1946. *Nature, Lond.*, **157,** 663.
Vago, E. E. and Barrow, R. F., 1946 a. *Proc. Phys. Soc.*, **58,** 538 ; 1946 b. *Ibid.*, **58,** 707.

THE LANDAU VELOCITY IN LIQUID HELIUM II

By D. V. GOGATE and P. D. PATHAK,
Physics Department, Baroda College, Baroda, India

MS. received 11 *February* 1946 ; *in revised form* 21 *November* 1946

ABSTRACT. A simple derivation of Landau's expression for the second velocity of sound (in addition to the usual velocity $\sqrt{(dp/d\rho)}$) in helium II is given from general physical principles.

§ 1. INTRODUCTION

The peculiar properties of liquid He II have been the subject of considerable experimental and theoretical research during the last few years. At 2·19° K. (the so called λ-point) helium shows a discontinuity in its specific heat, indicating a characteristic type of phase transition. Its viscosity decreases suddenly at the λ-point and the entropy difference between the liquid and the solid phase tends towards zero, showing that the liquid phase goes into a peculiar state below the λ-point. As is well known, heat flow in liquid He II is accompanied by a transfer of momentum (the thermo-mechanical effect). It has further been found that different experimental arrangements in the measurement of viscosity and thermal conductivity lead to different values, e.g. the viscosity vanishes when measured by the flow of liquid helium through a thin capillary or a narrow slit, while a non-zero value of viscosity is obtained when the latter is measured by the rotating-disc method. A similar state of things is found to exist with regard to the values of thermal conductivity. The systematic and thorough experimental investigation of the peculiar properties of He II is due to P. L. Kapitza (1938). He has shown the thermodynamic reversibility of the thermomechanical effect and has thus evolved a method of attaining temperatures approaching the absolute zero.

The unusual characteristics of He II led L. Tisza (1938) and F. London (1938) to suggest that liquid He II may be regarded as an ideal degenerate Bose-Einstein gas. Tisza suggested that the atoms found in the normal state (a state of zero energy) move through the liquid without friction, while London tried to explain the discontinuity in the specific heat at the λ-point, and the thermomechanical effect, by treating helium II as a Bose-Einstein degenerate system in which one fraction of the substance is distributed over the excited states in a way determined by the temperature, while the rest is condensed in the lowest energy level.

Recently Mendelssohn (1945) has advanced the hypothesis that both the superfluidity of He II and the superconductivity of electrons are caused by one and the same mechanism of "frictionless transport". The superfluid helium atoms, like the superconducting electrons, form an aggregate in momentum space of zero thermal energy (Z-state). The Z-particles have zero entropy but an appreciable zero-point energy. This theory is not based on any particular theoretical conception but is built up on the existing experimental evidence.

However, the most noteworthy theoretical study of He II is due to L. Landau (1941). In a most suggestive paper he has not only accounted for some of the previously known peculiar properties of He II, but has also predicted the existence of two velocities of sound in that liquid. This prediction has been experimentally confirmed by V. Peshkov (1944), who has obtained a value of about 20 m./sec. for the second (Landau) velocity in He II.

Landau has introduced the concept of a quantum liquid of which He II provides, so far, the only practical illustration. From the quantization of an arbitrary system of interacting particles (a liquid) he has advanced a system of hydrodynamic equations describing the macroscopic motion of liquid He II and has shown that two velocities of sound must exist in that liquid at non-zero temperatures. As the existence of a second velocity (in addition to the usual velocity $\sqrt{(dp/d\rho)}$) in He II is rather interesting, and has been experimentally verified, it may be worth while to discuss the existence of this abnormal velocity (which will be referred to hereafter as the Landau velocity) from general physical principles. This is done in this paper.

§ 2. ELEMENTARY DERIVATION OF THE EXPRESSION FOR THE LANDAU VELOCITY

Following Landau, we assume that liquid He II contains two types of fluid. The first type (n) is the ordinary (normal) type of fluid, having for its viscosity η nd entropy S their usual (non-zero) values. The second type of fluid (s) (which we may call a superfluid) possesses no viscosity ($\eta=0$) and no entropy ($S=0$). Let ρ_n, V_n and ρ_s, V_s denote the density and velocity of motion for the fluids n and s respectively.

Then the density ρ of liquid He II will be given by

$$\rho=\rho_n+\rho_s. \qquad ......(1)$$

The two types of fluid in He II do not resist the motion of each other, i.e. they possess independent motions. Thus for the flow of liquid He II we may write the equation

$$\rho V=\rho_n V_n+\rho_s V_s, \qquad ......(2)$$

where V denotes the average velocity of the fluids n and s.

We can contemplate a motion of the liquid when $V=0$. Then equation (2) will be reduced to

$$\rho_n V_n+\rho_s V_s=0, \qquad ......(3)$$

which means that there is no net transport of mass across any plane in the liquid.

Let us now consider the transport of entropy. For simplicity we consider the one-dimensional case. If we imagine a parallelopiped of unit cross-section and length Δx to be situated in the liquid with one pair of its opposite parallel faces vertical, the amount of entropy entering the parallelopiped per second at one face will be

$$\rho_n V_n S_n=\rho V_n S, \qquad ......(4)$$

where S is the entropy possessed by one gram of liquid He II.

The amount of entropy leaving the parallelopiped at the opposite parallel face per second will be

$$V_n\rho S+\frac{\partial}{\partial x}(V_n\rho S)\Delta x. \qquad ......(5)$$

The net transport of entropy per unit volume per second is given by

$$\frac{\partial}{\partial x}(\rho S V_n)=-\frac{\partial}{\partial t}(S\rho). \qquad ......(6)$$

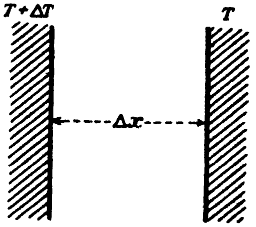
If we assume ρ to be sensibly constant and also take V_n to be small we obtain

$$\frac{\partial S}{\partial t}=-S\frac{\partial V_n}{\partial x}, \qquad ......(7)$$

neglecting the second order term $V_n\dfrac{\partial S}{\partial x}$.

Again the amount of entropy leaving the face at the temperature $T+\Delta T$ and reaching the opposite parallel face at temperature T will be given by

$$-\rho S V_n. \qquad ......(8)$$



Now we know from the second law of thermodynamics that $W/Q=\Delta T/T$, where W is the work done and Q the heat taken in at temperature T, or

$$\frac{\text{Work done}}{\text{Entropy absorbed by the system}}=\Delta T. \qquad ......(9)$$

We assume a reversible change in the case of liquid He II, and this is an essential condition for the existence of the abnormal velocity. Then, according to the second law of thermodynamics, the amount of entropy (which remains constant and which flows from one face to the other) multiplied by the difference of temperature between the two opposite parallel faces will be equal to the work done. Hence the entropy flow $\rho S V_n$ per second must produce work at the rate of $-(\rho S V_n)\Delta T$, and this will appear in the form of kinetic energy.

Now the kinetic energy per unit volume is given by $\frac{1}{2}[\rho_n V_n^2 + \rho_s V_s^2]$ and the rate of change of kinetic energy per unit volume is

$$\rho_n V_n \frac{\partial V_n}{\partial t} + \rho_s V_s \frac{\partial V_s}{\partial t} = \rho_n V_n \frac{\partial}{\partial t}(V_n - V_s) \text{ by equation (3)}$$

$$= \rho_n V_n \frac{\partial}{\partial t}\left(V_n \frac{\rho}{\rho_s}\right) = \rho\rho_n V_n \frac{\partial}{\partial t}\left(\frac{V_n}{\rho_s}\right)$$

$$= \rho\rho_n V_n \left[\frac{1}{\rho_s}\frac{\partial V_n}{\partial t} - \frac{V_n}{\rho_s^2}\frac{\partial \rho_s}{\partial t}\right]$$

$$= \frac{\rho\rho_n V_n}{\rho_s}\frac{\partial V_n}{\partial t}, \qquad \ldots\ldots(10)$$

neglecting the second-order term $\frac{V_n}{\rho_s^2}\frac{\partial \rho_s}{\partial t}$.

Hence the work done in volume Δx per second is

$$\frac{\rho\rho_n V_n}{\rho_s}\frac{\partial V_n}{\partial t}\Delta x, \qquad \ldots\ldots(11)$$

and by the second law of thermodynamics this must equal $-(\Delta T)\rho S V_n$.

Hence we have

$$S\frac{\partial T}{\partial x} = -\frac{\rho_n}{\rho_s}\cdot\frac{\partial V_n}{\partial t}. \qquad \ldots\ldots(12)$$

Again we know $dQ = cdT + p\,dv = Td\,S$, and when v is constant, $\partial T/\partial S = T/c$, where c is the specific heat.

Thus

$$\frac{\partial T}{\partial S} = \frac{T}{c} = \frac{\partial T}{\partial x}\frac{\partial x}{\partial S}$$

or

$$\frac{\partial T}{\partial x} = \frac{T}{c}\frac{\partial S}{\partial x}.$$

Hence from (12) we get

$$S\frac{T}{c}\cdot\frac{\partial S}{\partial x} = -\frac{\rho_n}{\rho_s}\cdot\frac{\partial V_n}{\partial T}$$

or

$$\frac{\partial V_n}{\partial T} = -\frac{\rho_s}{\rho_n}\cdot S\frac{T}{c}\cdot\frac{\partial S}{\partial x}. \qquad \ldots\ldots(13)$$

Differentiating (7) with respect to t and (13) with respect to x, we get

$$\frac{\partial^2 S}{\partial t^2} = \frac{\rho_s}{\rho_n} \cdot \frac{S^2 T}{c} \cdot \frac{\partial^2 S}{\partial x^2} \qquad \ldots\ldots(14)$$

or

$$u^2 = \frac{\rho_s}{\rho_n} \cdot \frac{S^2 T}{c} \qquad \ldots\ldots(15)$$

where u (or u_2) is the second velocity (Landau velocity) of sound.

REFERENCES

Kapitza, P. L., 1938. *Nature, Lond.*, **141**, 74 ; 1940. *J. Phys. U.S.S.R.*, **4**, 181 ; 1941. *Phys. Rev.*, **60**, 354.
Landau, L., 1941. *Phys. Rev.*, **60**, 358 ; *J. Phys. U.S.S.R.*, **5**, 71 ; 1944. *J. Phys. U.S.S.R.*, **8**, 1.
London, F., 1938. *Nature, Lond.*, **141**, 643 ; 1939. *J. Phys. Chem.*, **43**, 49.
Mendelssohn, K., 1945. *Proc. Phys. Soc.*, **57**, 371.
Peshkov, V., 1944. *J. Phys., U.S.S.R.*, **8**, 381.
Tisza, L., 1938. *Nature, Lond.*, **141**, 913.

THE RADIO DETECTION OF METEOR TRAILS AND ALLIED PHENOMENA

By Sir EDWARD APPLETON, F.R.S. and R. NAISMITH

Read 31 *January* 1947; *MS. received* 21 *February* 1947

ABSTRACT. The results of a series of radio sounding observations on (*a*) transient bursts of atmospheric ionization, and (*b*) abnormal or sporadic E-layer ionization, are described and discussed. Special observations during the recent Giacobinid meteor shower of 10 October 1946 confirm that the ionization bursts which can be observed at all hours of the day are due to sporadic meteors. A statistical study of abnormal or sporadic E-layer ionization gives strong support to the view that, in temperate latitudes, this phenomenon is also largely ascribable to the same cause.

§ 1. INTRODUCTION

It has been known for many years that the ionization in the E-layer of the ionosphere is subject to sporadic changes. Such changes are due to the incidence of irregular ionizing agencies which are additional to the normal ultra-violet radiation from the sun. In 1935 (Appleton and Naismith, 1935), for example, evidence was presented which indicated the separate existence and unlike behaviour of abnormal and normal E-layer ionization and by which it was possible to demonstrate that the normal ionization due to solar ultra-violet light followed a very regular law as regards its dependence on solar zenith distance; and that the great variability of E-layer ionization as a whole was due mainly to the variability of the additional agency responsible for abnormal ionization.

In this paper we describe the results of a series of radio investigations, carried out in recent years, which shed some light on the origin of the irregular ionization in the lower ionosphere. Using the ordinary methods of radio sounding by pulses, two types of this sporadic ionization have been recognized. It may well be, as suggested below, that these have a common origin, but the two manifestations of it are sufficiently dissimilar in character for us to treat them separately.

We shall call the two types of ionization:—Type (A), E-layer bursts of ionization, and Type (B), Abnormal or sporadic E-layer ionization; and discuss them in the order given.

§2. E-LAYER IONIZATION BURSTS (TYPE A)

During the International Polar Year 1932–3 a transient type of echo was observed (Appleton, Naismith and Ingram, 1937) at about the level of the E-layer. This echo was found to last only a second or two, and occurred both by night and day. The fact that there was no major difference in incidence between night and day was considered to exclude solar ultra-violet light as the responsible agency; and it was pointed out that, possibly, meteors were responsible for these ionization bursts, since Skellet (1935) had observed that major increases of Abnormal E ionization (i.e. Type (B) above) occurred at night when meteors were observed to pass overhead.

A fairly extensive study of E-layer bursts of ionization was later made by Appleton and Piddington (1938), at Cambridge, who used a 2 to 3 kilowatt pulse sender on 8·8 Mc./s. coupled to a dipole aerial which radiated vertically upwards. These authors made a study of the distribution of echo occurrence in range and found that this distribution was very similar by day and by night. They also measured the average equivalent reflection coefficient of ionization bursts, for the particular frequency employed. The same type of observations was continued at Cambridge by Mohanty (1938), in an extensive series of experiments, who showed that the reflections from ionization bursts continued to be appreciable even when the frequency used was increased to 16 Mc./s., although the effective reflection coefficient was found to be less as the frequency was increased. Mohanty further came to the conclusion that the ionization bursts (Type A) were the trials of meteors.

§3. ABNORMAL OR SPORADIC E-LAYER IONIZATION (TYPE B)

Using radio frequencies slightly in excess of the critical frequency of the normal E-layer, it is usually possible to detect reflections from more intensely ionized clouds or strata embedded in the normal E-layer. Such clouds or strata are called *abnormal or sporadic E-layer ionization.* They differ from ionization bursts in that they are more persistent in time and, in daytime, appear at a very limited range of heights. They are particularly in evidence in summer and in the daytime. On occasion, and most frequently in summer, this ionization may reach such high densities that its critical penetration frequency exceeds that of the F_2-layer. We have referred to such manifestations as "Intense E-ionization" (Appleton, Naismith and Ingram, 1939; Appleton and Naismith, 1940). Under conditions of marked abnormal E-layer ionization, the upper limit of frequency usuable for long-distance communication may be abnormally high, in that the usual restriction by way of F_2-layer electron limitation is not operative.

Abnormal or sporadic E-layer ionization, at any rate in the daytime, is usually situated somewhere near the level of the maximum electron density of the normal E-layer. Reflections from it are also fairly markedly frequency-dependent in that, as the radio frequency employed is increased, the intensity of reflection falls off fairly rapidly within a definite frequency range. Before echoes become inappreciable there is, however, often a frequency interval of partial reflection in which echoes are simultaneously observed from both the abnormal E-layer and from the F-layer.

§ 4. OBJECT OF THE PRESENT SERIES OF EXPERIMENTS

The present series of observations was undertaken with the general object of comparing the incidence of Type A and Type B ionization in the E-layer and, in particular, of testing how far the meteor theory accounted for Type A.

For the experiments on ionization bursts, a high-power pulse sender and a receiver were kindly lent to us by the Air Ministry. In view of the fact that the earlier observations had been conducted on the lower range of frequencies it was decided to use a higher radio-frequency of 27 Mc./s. These observations were conducted on a flat site at the Radio Research Station where an aerial system designed to radiate vertically upwards was erected for us by 60 Group of the Royal Air Force. Pulses of 15 μs. and of a repetition rate of 50 per second were used and the usual cathode ray oscillograph display of ground and echo pulses was employed.

For the study of abnormal or sporadic E-layer ionization we were able to rely on results from the daily soundings of the ionosphere conducted at the Slough Radio Research Station, which form part of the continuous series of measurements of ionospheric layer densities which we have made since 1931, when the critical-frequency method was first introduced (Appleton, 1931; Appleton and Naismith, 1932).

For convenience we describe the observations on the two types of abnormal ionization in two sections below dealing respectively with Type A and Type B.

§ 5. EXPERIMENTAL RESULTS: PART I

In this section we deal with the observations on ionization bursts using a radio exploring frequency of 27 Mc./s. By employing photographic registration we have endeavoured to eliminate the subjective feature of the earlier observations made in Cambridge. We have also tried to keep the experimental conditions as constant as possible over a period of over two years, so that the seasonal variation of ionization bursts could be examined, Supplementary eye observations have, however, also been made. Special series of observations have also been made at times of expected meteor showers, notably on 10 October 1946, on the occasion when the Giacobini-Zinner comet approached the earth.

Perhaps the most remarkable result of the continuous series of observations is the seasonal variation of ionization-burst occurrence at noon. In figure 1 are shown the monthly means of the number of bursts per hour at noon for the period 1944–1946. It will be seen that, although summer values are higher than the winter values, the curve is not symmetrical about the summer solstice.

(In the same figure are shown, for comparison, the monthly means of the frequencies of occurrence of noon sporadic E-layer ionization on 4 Mc./s. This comparison will be referred to below.)

In our study of the diurnal variation of ionization bursts, attention has been specially directed to nocturnal events, so that a comparison could be made with the results of visual meteor observations. In figure 2 is illustrated the nocturnal variation of ionization-burst frequency, from which it will be seen that the rate of ionization-burst detection is greater after midnight than before. In figure 3 is shown the nocturnal variation of ionization-burst duration, showing that the duration of these transient echoes is also greater after midnight than before. Since we know that, for constant experimental conditions, echo duration usually increases with echo intensity, we interpret the latter result as indicating that stronger reflections from ionization bursts are obtained in the second half of the night than in the first.

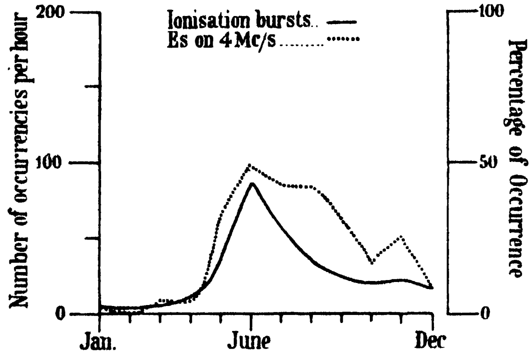


Figure 1. Seasonal variation of frequency of ionization bursts and of frequency of occurrence of sporadic E-layer ionization (4 Mc./s.) in daytime. (Average 1944–6.)

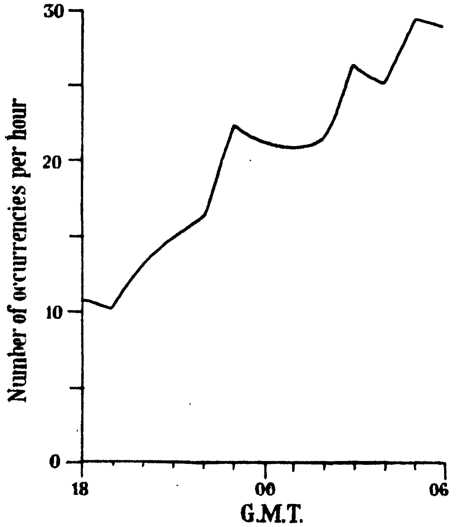


Figure 2. Nocturnal variation of ionization burst frequency.

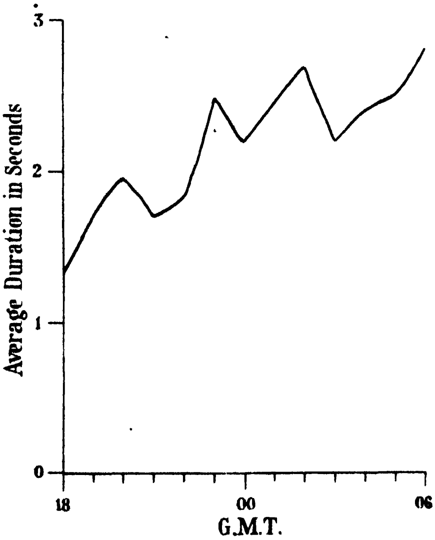


Figure 3. Nocturnal variation of ionization burst duration.

In figure 4 are shown typical diurnal curves of ionization-burst frequency in summer and winter. A remarkable difference is noted between these two curves. The night-time variation is found to be fairly constant in character, but the day-time portion alters with the season of the year. For example, it is found that there is a characteristic minimum about 1800 G.M.T. throughout the year. It is also found that the number of bursts increases from that time onwards until about 0500 G.M.T. in the winter, the average number of bursts in the second half of the night being about twice the value during the first half. Under these winter conditions the early morning maximum is the maximum of the day. On the other hand, in summer, the increase in echo occurrence in the second half of the night continues after dawn so that the maximum of the day is reached just before, or at, local noon.

Before attempting to interpret the above results in terms of the meteor theory of origin, we describe first what we think is the most weighty evidence for this theory of which we are aware. Ionospheric literature contains many examples of simultaneous experimental observations on abnormal E-layer ionization and visual observations on meteors, and many workers have concluded that such ionization was increased at such periods of meteoric activity (Skellet, 1931; Mitra, Syam and Ghosh, 1934; Bhar, 1937; Pierce, 1938). More recently it has been considered desirable to concentrate such radio observations, not on abnormal E-ionization but on ionization bursts, and Hey and Stewart (1946) have recently shown a correlation between ionization bursts and meteor showers. By making daily observations with a radio beam directed vertically upwards, they found marked peaks in the frequency of occurrence of ionization bursts coinciding with major meteor showers. Further, it was found that stations with inclined radio beams gave different diurnal variations when set on different bearings; by making the assumption that the most favourable direction of observation was at right angles to the meteor train, they were in certain cases able to derive from these results the approximate positions of the meteor radiants.

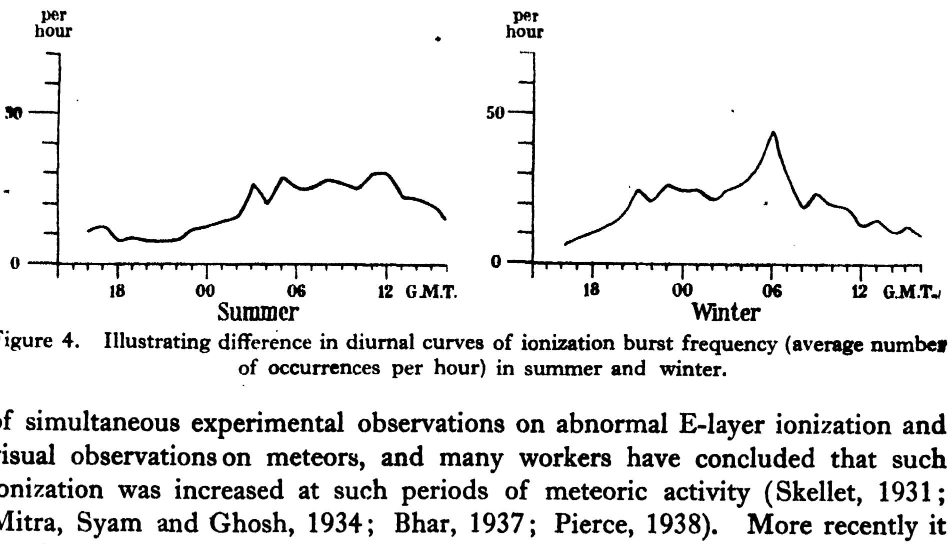


Figure 4. Illustrating difference in diurnal curves of ionization burst frequency (average number of occurrences per hour) in summer and winter.

We have found similar correlations between increases of ionization-burst frequency and the incidence of meteor showers, but not one of them is nearly as striking as that which occurred on 10 October 1946 on the occasion of the shower associated with the Giacobini-Zinner comet. The results on that occasion are shown in figure 5, together with the results for the control days, for the same night hours, about the same period. It will be seen from these curves that on the control days, there was the usual increase of ionization-bursts detected during the night; but on the occasion of the meteor shower there was an enormous increase during the period 0300 to 0400 G.M.T. For the curve shown in figure 5 the burst-frequency rate was calculated by taking rather long time intervals of one hour. By taking the much smaller time interval of one minute, the detail of the variation of rate is disclosed. In figure 6 is therefore shown how the rate of bursts, estimated in this way, varied with time throughout the shower.

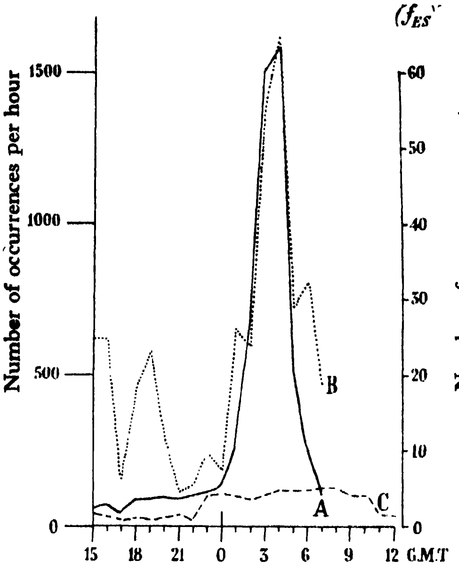


Figure 5. Showing variations of (A) ionization burst frequency, and (B) sporadic E-layer ionization during the night of the Giacobinid meteor shower. (A corresponding curve of ionization burst frequency (C) for a normal night is also shown.)

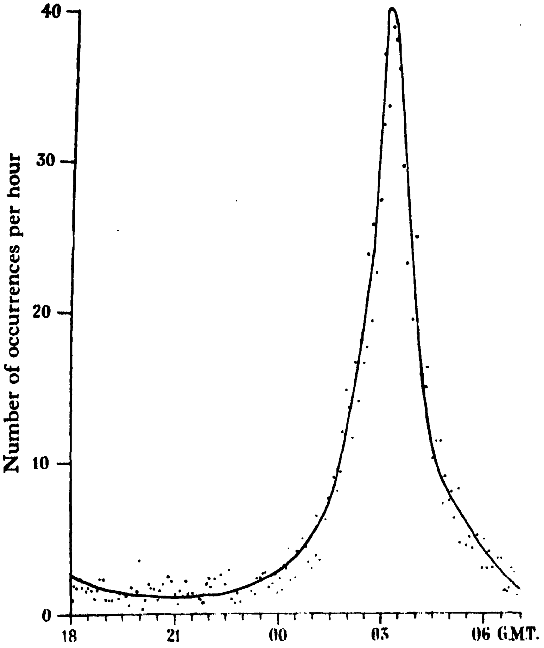


Figure 6. Showing detailed variation of ionization burst frequency during the Giacobinid meteor shower of 9/10 October 1947.

In plates A and B are shown some sample records for various times during the night of 9/10 October. These clearly show the remarkable display of ionization bursts which accompanied the meteor shower.

§6. DISCUSSION OF EXPERIMENTAL RESULTS: PART I (*continued*)

We now turn to discuss the results we have obtained, so far, on the subject of ionization bursts. There can be no doubt from the experiments of 10 October that meteor trails do give echoes of the type we have called ionization bursts.

It therefore remains to be considered whether all such bursts can be attributed to meteors. In this connection it is important to have regard to

(*a*) the diurnal and seasonal incidence of sporadic meteors;
(*b*) the exceptional influence of particular meteor showers; and
(*c*) the possible diurnal and seasonal variation of the response of the atmospheric medium.

In this connection it is well to recall that visual observations of meteors, are, of course, only possible during the hours of night-time; and the moon, when nearly full, makes such observations well nigh impossible. Such observations show, on the average, both a diurnal and seasonal variation in the number of visible meteors. The average hourly number after midnight is about double the average number before midnight. This is due to the fact that, in the evening, the meteors must overtake the earth, whereas in the morning hours we see both the meteors which the earth overtakes and those coming to meet it. Also meteors are twice as numerous from July to January as from January to July. These remarks refer to sporadic meteors which are not associated with any recognized shower.

In general it will be seen that our results fit in with what we should expect according to the meteor theory in that, during the night, the number of ionization bursts after midnight is approximately twice what it is before midnight. Concerning the annual variation it will be seen that the maximum in the curve of ionization bursts does not occur in the second half of the year, but in summer. The September values are, however, definitely higher than the March values. It appears unlikely that the curve in figure 1 indicates the true seasonal variation of sporadic meteors at noon, and we suggest that the variation disclosed indicates some form of solar control of the "response" of the atmospheric medium. For example, let us suppose that the "response" of the medium to the ionizing particles varies in some such manner as does the ionization in the E-layer, there being a maximum in summer and a minimum in winter. If, then, the incidence of the ionizing agency were greater in the second half of the year than in the first, the general shape of the resultant curve of detected ionization bursts might be expected to be as shown in figure 1.

§ 7. EXPERIMENTAL RESULTS: PART II

As mentioned above, we have relied for our results on abnormal or sporadic E-layer ionization on the Slough daily (h', f) ionospheric records. We have found that, generally, these records for the years 1945 and 1946 show results similar to those we have previously obtained and described (Appleton, Naismith and Ingram, 1939; Appleton and Naismith, 1940). Sporadic E-layer ionization is found to be greater in summer than in winter and greater by day than by night. But, by more accurate measurements, we have proved, more definitely than before, that abnormal E-layer ionization is controlled, at any rate in the daytime, in its magnitude and level of occurrence, by the normal E-layer. For example, in figure 7 is shown the diurnal variation of height of abnormal E-layer reflections, from which it will be seen that the level varies in exactly the same way as does the level of maximum E-layer ionization during the daytime. Moreover, we have found that the seasonal variation of average height at noon varies as we should expect if it is associated with the level of the normal E-layer. This is illustrated in figure 8.

It has been shown by Skellet and others that there is a temporary increase in abnormal E-layer ionization during periods of meteoric showers, and our own observations, illustrated in figure 5, where both ionization bursts and abnormal E-layer ionization are plotted for the night of 10 October, illustrate this same effect in a most striking way. But abnormal or sporadic E-layer ionization occurs at times other than those of meteor showers, and the real question at issue is whether or not we can consider meteors as one of the major sources of sporadic E-layer ionization, bearing in mind its very striking diurnal and seasonal characteristics which indicate some form of solar control.

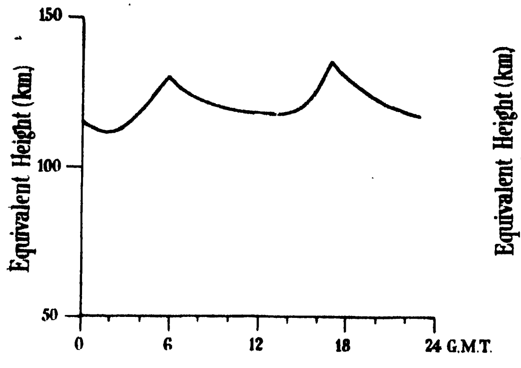


Figure 7. Showing diurnal variation of the average height of sporadic E-layer ionization.

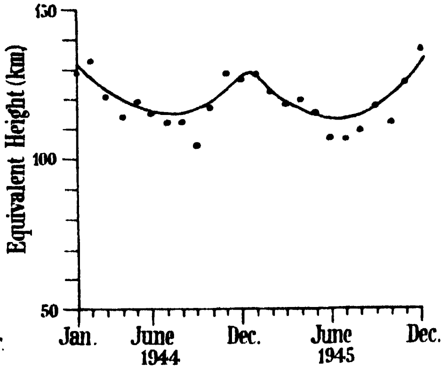


Figure 8. Showing seasonal variation of the average height of sporadic E-layer ionization at noon.

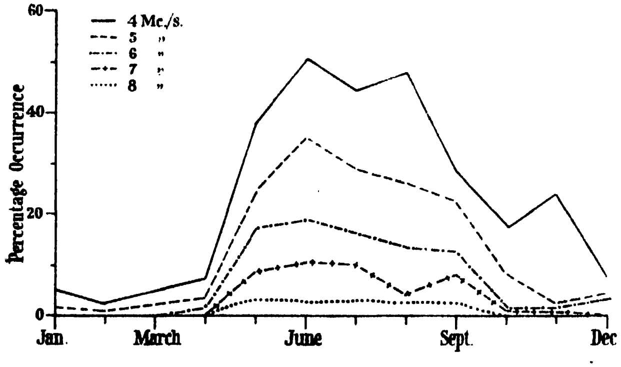


Figure 9. Percentage occurrence of sporadic E-layer reflections at noon at Slough on fixed radio frequencies (average for 1943–6).

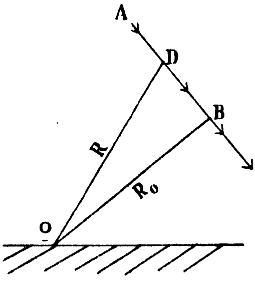


Figure 10.

The results described above, however, indicate precisely the same kind of solar control for ionization bursts, which we now can definitely attribute to meteors. The parallelism between the two phenomena of sporadic E-layer ionization and meteor-trail incidence is well illustrated in figure 1. The annual variation of sporadic E-layer ionization at noon is, perhaps, better illustrated in figure 9 where the percentage occurrence of sporadic E-layer reflections at noon at Slough for various frequencies is shown. The asymmetry of these curves relative to midsummer is clearly marked.

We therefore suggest that there are strong grounds for accepting sporadic meteors as being responsible in substantial measure for sporadic E-layer ionization in temperate latitudes. It should, however, be noted that, taking the earth as a whole, it has already become clear that there must be at least two sources of sporadic E-layer ionization. In the course of the International Polar Year investigations (Appleton, Naismith and Ingram, 1939) it was found that there was, in high latitudes, a high positive correlation between sporadic E-layer ionization and magnetic activity. We can therefore conclude that near the auroral belt, sporadic E-layer ionization can, to a large extent, be attributed to the cosmic agency (e.g. charged solar particles) responsible for the causation of magnetic storms.

On the other hand, at the lower latitude of Slough (51½° N.) there is no marked connection between sporadic E-layer ionization and magnetic activity, and the diurnal variation of such ionization has entirely different characteristics from those exhibited in high latitudes. A different originating agency is therefore suspected.

§ 8. SUMMARY OF CONCLUSIONS

It may be convenient to summarize briefly the results and conclusions emerging from the above discussion.

(*a*) Earlier conclusions, that transient radio echoes from the level of the E-layer are due to reflections from the ionization trails of meteors, are confirmed. The effective scattering area of such trails falls off as the sounding radio-frequency is increased. Thus, if high radio frequencies are used, only the denser trails (e.g. in meteor showers) are detected. With lower radio frequencies, the general incidence of sporadic meteors may be studied.

(*b*) There is a marked seasonal variation of the number of sporadic meteors recorded per hour at noon, the number being a maximum in summer and a minimum in winter. The curve illustrating this relation is not, however, symmetrical about midsummer, for it is found that more meteors are recorded in the second half of the year than in the first. It is not considered that this curve necessarily gives a correct indication of the seasonal variation of the actual incidence of meteors at noon, and a possible seasonal variation of the "response" of the atmospheric medium to the ionizing meteroric particles is suggested.

(*c*) Average curves of the diurnal variation of recorded ionized meteor trails fit in with the known features of the incidence of sporadic meteors so far as night-time conditions are concerned. The daytime portion of the diurnal curve of meteors recorded by radio methods is, however, found to alter with the season, most probably because the daytime "response" of the medium to the ionizing meteoric particles (referred to in (*b*) above) is greater in summer than in winter.

(*d*) During the Giacobinid meteor shower of 10 October 1946, a marked increase of sporadic E-layer ionization was found to accompany the very marked increase of ionized meteor trails recorded. This, and the striking parallelism between the daytime incidence of sporadic E-layer ionization and meteor ionization bursts, have led the authors to advance the theory that the fine dust of ionizing meteors is mainly responsible for the production of sporadic E-layer ionization in temperate latitudes.

(*e*) Measurements of the heights of occurrence of meteor ionization bursts and sporadic E-layer ionization indicate that, at any rate in the daytime, the former occur at a slightly lower level than does the latter. The atmosphere is, therefore, pictured as being bombarded with meteors and especially with dust particles too small to produce visible meteor trails. As these particles strike through the E-layer they cause additional ionization which, in the daytime, is detected by the usual methods of ionospheric sounding at the level of maximum E-layer ionization. During the night, sporadic E-layer ionization is detectable over a greater range of heights.

APPENDIX

A note on whistling meteors

Some years ago Chamanlal and Venkataraman (*Electrotechnics*, No. 14, Nov. 1941) described some very interesting experiments in which they observed a radio Doppler effect due to meteors entering the earth's atmosphere, which was exhibited in the form of whistles heard in the telephone of a radio set tuned to an unmodulated short-wave radio station. The observed whistles were found to descend in pitch, the audible note usually continuing to zero frequency, though on occasion the note ceased before zero frequency was reached. The explanation of the phenomenon given by Chamanlal and Venkataraman was that the whistles were heterodyne notes caused by the interference of the direct radio wave from the sending station and the reflected wave from the rapidly moving ionized surface associated with the meteor. The same authors further stated that "the descending pitch of the heterodyne beat note is the result of the moving reflecting surface being rapidly retarded in velocity, since it is evident that the beat note will reduce to zero frequency if the velocity of the reflecting surface becomes zero."

In our own observations on meteor reflections we have often observed whistles of falling frequency of this kind, and, through the kindness of Mr. L. W. Hayes of the B.B.C., we have been furnished with observations on the same subject made at the B.B.C. Listening Station at Tatsfield.

As a result of our attempt to correlate the results obtained by the two methods of meteor detection, by pulses and by Doppler whistles, we have been led to question the validity of the explanation of the varying pitch of the whistles, due to Chamanlal and Venkataraman, and quoted above. We may illustrate the matter with reference to figure 10 where the simplest type of experimental arrangement is depicted and where O may be taken as the site of both sending and receiving stations (pulse and C.W.).

Let us suppose that the meteor trail is ADBC and that the meteor velocity is v. Further, let t be the time when the meteor is at D, and thus at range R; and let t_0 be the time when the meteor is at B, the minimum distance. Let this minimum distance be R_0. During the period when the meteor is travelling from A to B we have

$$R^2 = R_0^2 + v^2(t_0 - t)^2, \qquad \ldots\ldots(1)$$

and thus

$$\frac{dR}{dt} = -v\left(1 - \frac{R_0^2}{R^2}\right). \qquad \ldots\ldots(2)$$

For the observations on continuous waves, the beat frequency observed at 0, due to the interference between the direct wave and the wave reflected from the meteor trail, is given by

$$f=\frac{2}{\lambda}\frac{dR}{dt}=-\frac{2v}{\lambda}\left(1-\frac{R_0^2}{R^2}\right)^{\frac{1}{2}}. \qquad \ldots\ldots(3)$$

Thus as the meteor particle travels from A to B the value of f should fall and this is found to be the case. The frequency will become zero when R is equal to R_0. We thus see that the fall of the whistle frequency can be accounted for without assuming retardation of the meteor itself. The insertion of typical values in the right-hand equation of (3) gives values of f of the order actually measured.

Now Pierce (1938) has emphasized the fact that the strongest reflection may be expected from a meteor trail which is broadside on to the exploring radio station. In other words, the strongest reflection received back at 0 will come from B on the meteor trail. Experimental evidence supporting this has been given by Hey and Stewart (1946). In our own experiments using radio pulses, we believe that it is usually the distance R_0 which is measured, since the value of the range remains fairly constant during the period when the echo is received.

(It may be of more than theoretical interest to note that, if a record of whistle frequency with time could be obtained, together with a pulse determination of the range R_0 for the same meteor, it would be possible to find the meteor velocity v. For, from (3), we have

$$R=R_0+\frac{\lambda}{2}\int_{t_0}^{t} f\,.\,dt, \qquad \ldots\ldots(4)$$

so that the value of R could be inserted in the right-hand side of (3) and the value of v found.)

It will be seen that, according to the explanation of the connection between meteor whistles and meteor pulses echoes given above, the former phenomenon occurs earlier than the latter. Now, it has been frequently reported by observers of meteor whistles, listening within the "skip" distance to a near-by short-wave sending station, that the whistles are succeeded by a burst of signal strength. We think that the burst of signal strength corresponds to the arrival of the meteor head at the broadside-on position B (see figure 10) and so corresponds in time with the reception of the strong pulse echo.

ACKNOWLEDGMENT

The work described above was carried out as part of the programme of the Radio Research Board of the Department of Scientific and Industrial Research.

REFERENCES

APPLETON, 1931. *Nature, Lond.*, **127**, 197.
APPLETON and NAISMITH, 1932. *Proc. Roy. Soc.*, A, **137**, 36.
APPLETON and NAISMITH, 1935. *Proc. Roy. Soc.*, A, **150**, 685.
APPLETON and NAISMITH, 1940. *Proc. Phys. Soc.*, **52**, 402.
APPLETON and PIDDINGTON, 1938. *Proc. Roy. Soc.*, A, **164**, 467.
APPLETON, NAISMITH and INGRAM, 1937. *Phil. Trans. Roy. Soc.*, A, **236**, 191.
APPLETON, NAISMITH and INGRAM, 1939. *Proc. Phys. Soc.*, **51**, 81.

BHAR, 1937. *Nature, Lond.*, **139**, 470.
HEY and STEWART, 1946. *Nature, Lond.*, **158**, 481.
MITRA, SYAM and GHOSH, 1934. *Nature, Lond.*, **133**, 533.
MOHANTY, 1938. Ph.D. Thesis (Cambridge).
PIERCE, 1938. *Proc. Inst. Rad. Engrs.*, **26**, 892.
SKELLET, 1931. *Phys. Rev.*, **37**, 1668.
SKELLET, 1935. *Proc. Inst. Rad. Engrs.*, **23**, 132.

DISCUSSION

Mr. G. R. M. GARRATT. By the courtesy of Mr. Cecil Goyder, Chief Engineer of All India Radio, I was privileged a few months ago to observe the phenomenon of " whistling meteors " in New Delhi and to meet the two members of his staff, Messrs. Chamanlal and Venkataraman, who share the credit for this very interesting discovery.

I believe it was during the winter of 1940/41 that this effect was first observed. Whistles were heard while monitoring the short-wave transmitters of All India Radio which could not be explained as ordinary heterodynes. The whistles were of short duration, normally from about 2 to 3 seconds; each commenced as a note of about 3000 cycles, fell rapidly in pitch to zero or sometimes died away before reaching zero, and only in rare cases did the whistle reappear as an ascending note on the other side of zero.

It was eventually deduced that these unusual characteristics could only be explained on the assumption that they were due to a Doppler effect resulting from the interference between the weak ground waves received direct from the transmitter about 9 miles away and waves reflected from some very rapidly moving surface. The velocity of the reflecting surface towards the observer was shown to be of the order of 40–80 km./sec. and it seemed that the only likely source of such high velocities would be a meteor producing ionization in the earth's upper atmosphere. This assumption was easily confirmed in the clear sky of Northern India where " shooting stars " can be seen on almost any night of the year by establishing direct correlation between the occurrence of a whistle and the arrival of a visible meteor.

Although the whistles can be detected by any sensitive receiver capable of receiving broadcasts on the short wave bands, there are a number of conditions which must be satisfied before success will be achieved. Firstly, it is necessary to have a high power transmitter—preferably not less than 50 kw.—radiating an unmodulated signal on a frequency of the order of 5–15 Mc./s. The receiver must be situated within the skip distance so that no reflections are received from the E- and F-layers, yet it must be so located that the direct ground wave is a weak one. If the ground wave received is too strong, the normal operation of the A.V.C. circuits cuts down the amplification of the receiver and prevents detection of the very weak reflected waves. There is thus an optimum strength of ground wave which, with a normal receiver, is of the order of 70 to 100 mv. It has been estimated that the reflected signal from the meteor may be as small as 0·5 mv. although in the case of a meteor passing near the zenith, the energy received may be several hundred times as great as this.

Even in the clear sky of Northern India, only a small number of meteors are visible to the eye compared with the number which can be heard. The ratio is probably of the order of 100 : 1—though any meteor which is visible can be relied on to produce a prominent whistle.

These effects can be heard on almost any night of the year—the greatest number being heard about 4.0 a.m. The number to be heard on an ordinary night, however, may be only of the order of 4–6 per minute as compared with an almost continuous performance at times of meteoric showers.

It has been stated that the " whistle " is invariably characterized by a falling note, which has been attributed to the loss of velocity of the meteor as it is retarded in the earth's atmosphere. Although it is true that the great majority of whistles exhibit a falling note (which often dies away before reaching zero frequency), a small number yield a note which falls in pitch, passes through zero, rises again and then dies away at about 1 to 2000 cycles. A very small number seem to show only a rising note.

I think it is clear that these effects are due to a combination of circumstances. A Doppler effect or frequency shift arises from the component of velocity towards the observer and if the reflecting surface had a fixed component of velocity towards the observer, a beat note of

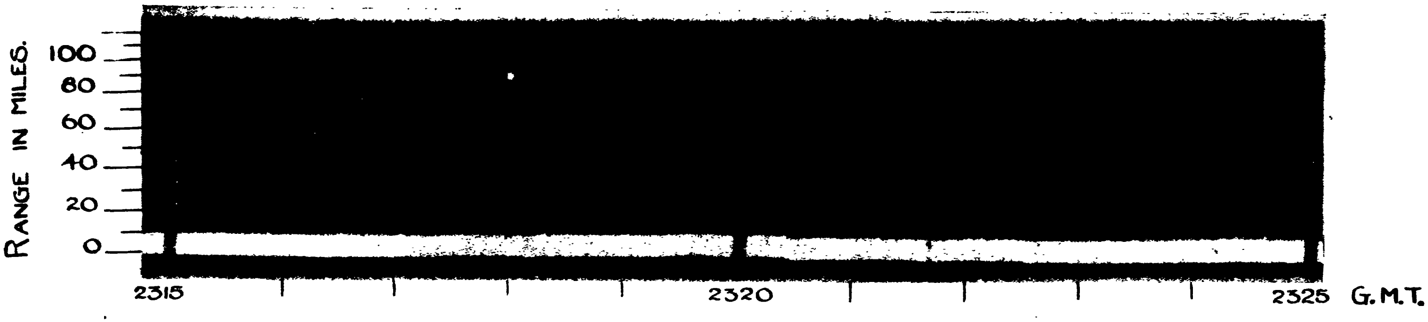


Plate A.

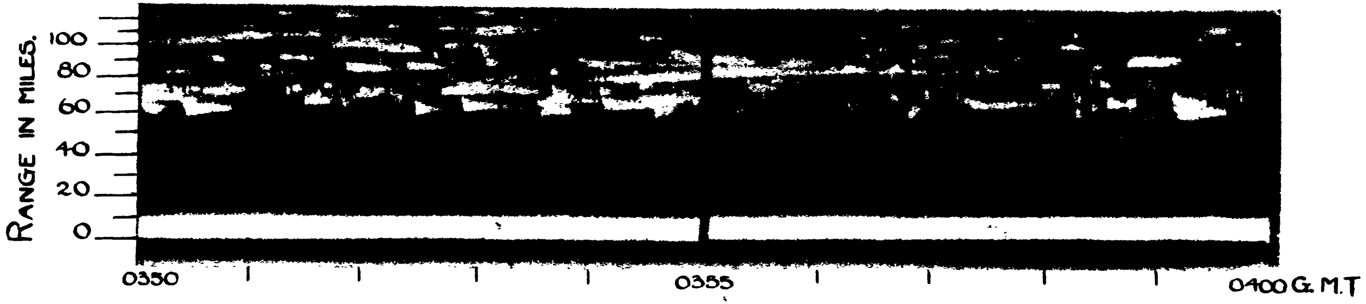


Plate B.

Ionization trails recorded at Radio Research Station, Slough. Plate A on 9.10.46; plate B at the height of the Giacobinid meteor shower on 10.10.46.

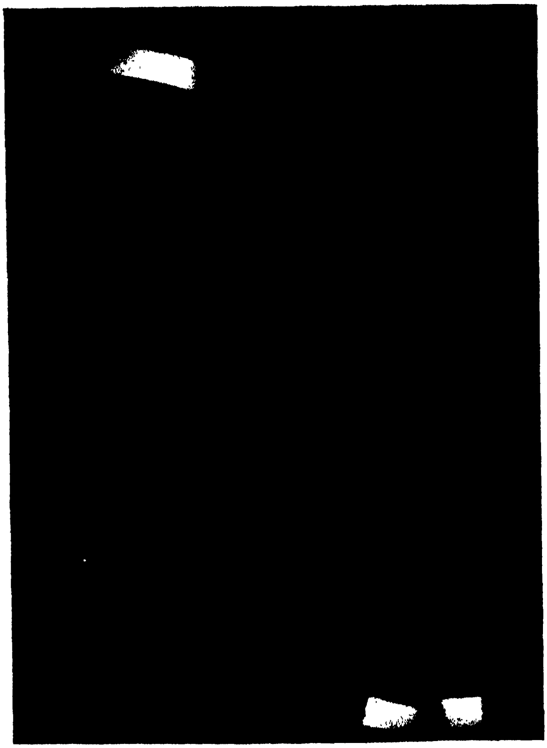


Figure 1. Illustration of the type of trace obtained (positive).

constant frequency would be heard. A falling note can be produced either by a retardation of the meteor in the upper atmosphere, in which case the component of velocity towards the observer also diminishes, or it can be produced by a meteor travelling at constant velocity in such a direction that its track lies tangentially with respect to the observer within the somewhat restricted region in which reflected echoes can be detected. For example, a meteor travelling horizontally and passing directly through the observer's zenith has a component of velocity towards the observer which diminishes to zero as it passes overhead and then increases away from him. Such a meteor travelling at constant velocity will give rise to a descending note as it approaches the zenith and an ascending note as it recedes.

In practice, the beat note heard must be almost invariably a combination of both effects, the retardation of the meteor in the atmosphere and the changing component of velocity with respect to the observer. Since meteors are retarded very rapidly in the earth's atmosphere, however, it is generally only those travelling on an approximately horizontal course at extreme altitudes where there is little retardation which have any appreciable velocity left after passing the zenith to produce a rising note at the receiver.

The mechanism by which a meteor gives rise to a beat note of rising frequency only, without a preceding fall through zero frequency, is not very clear. It seems possible, however, that it is not unconnected with the fact that the transmitter and receiver are situated some miles apart and the response received may therefore depend to some extent on the direction of the reflecting surface in relation to that of the transmitter.

Finally, it may be useful to note that if the meteor has a component of velocity towards the observer, the heterodyne whistle appears on the higher-frequency side of the ground wave. If the meteor is moving away, however, the whistle appears on the lower side of the ground transmission, and if it is specially desired to observe the somewhat rare cases where a rising note is exhibited, it is advisable to detune the receiver slightly on the low-frequency side, since the whistle in these cases is always a weak one and may otherwise be missed altogether.

THE SHORT-PERIOD TIME VARIATION OF THE LUMINESCENCE OF A ZINC SULPHIDE PHOSPHOR UNDER ULTRA-VIOLET EXCITATION

By MARY P. LORD,* A. L. G. REES† and M. E. WISE,
Philips' Lamps Ltd., Mitcham, Surrey
* Now at Imperial College, London
† Now in Melbourne, Australia

MS. received 20 *February* 1946 ; *in revised form* 18 *October* 1946

ABSTRACT. A photographic method, in some respects new, is described for measuring in detail the luminescent-intensity time relationships of a powdered phosphor. A zinc sulphide with silver and copper impurities was irradiated at 20, 40, 60 and 80° c. with various intensities of $\lambda 3650$: the results are mainly on the blue phosphorescent intensity over some 300 milliseconds from the end of irradiation.

The observations are found to correspond quantitatively to a bimolecular law with two types of activating centre. On this assumption four independent constants are deduced mathematically for each curve, two per centre. They are interpreted in terms of the initial concentrations of ionized centres, and the recombination coefficient for each kind of centre with free electrons.

Using these constants, the ratio of the contributions from the two kinds of centre to the luminescent intensity at the beginning of the decay was deduced. From this ratio and the spectral distribution of the fluorescent emission, the centre with the faster decay was

identified as silver and the other as interstitial zinc. The very slowly decaying green band present in the luminescent emission is ascribed to the presence of copper centres.

The constants are found to change with temperature and with exciting intensity in a way which cannot be explained on the simple theory. These variations can, however, be accounted for by assuming that there are trapping centres closely associated with each "excited" activating centre. It is shown that the traps associated with the copper centres are of major importance, although the concentration of copper impurity is relatively small.

As functions of exciting intensity the constants also have periodic terms superimposed on the main changes. An attempt is made to relate this qualitatively to the equations for the build-up of luminescence when more than one type of activating centre is present.

The effect of the thickness of the layer of powder on the simplest type of two-centre bimolecular decay law and the relation, in outline, of our interpretation to the work of other observers are also considered.

§ 1. INTRODUCTION

THE mechanism of the luminescence of inorganic solids has been the subject of much theoretical and practical investigation for many years, but in spite of the stimulus to this research provided by recent technological applications, the problem has remained very far from solution.

An experimental method has been evolved which permits observations to be made at time intervals of the order of 4 milliseconds over a range of 400 milliseconds from the beginning of the decay. Thus the results may be more detailed than any previously published over the range investigated.

§ 2. EXPERIMENTAL

The principle of the method used has been applied in many ways by earlier workers. A well-defined beam of monochromatic ultra-violet radiation strikes normally a layer of the luminescent material spread evenly on a glass disc that can be rotated uniformly at an appropriate speed about its axis. The light from the phosphor is recorded photographically; and, from the density of blackening at various points of the trace and the plate characteristics, the intensity of the phosphorescent light may be obtained.

The whole apparatus was mounted on a substantial base. Exciting radiation was obtained from the quartz inner tube of an MBF/V. 80-watt mercury-discharge lamp operated on d.c., the power consumed being controlled by a rheostat and measured on a Hartmann and Braun wattmeter. The radiation from the lamp was passed through a Wratten 18A filter to ensure that approximately monochromatic radiation of wave-length 3650 A. was incident on the phosphor. The beam was collimated as far as possible by passing through a series of limiting apertures, the final one being just clear of the phosphor and shaped as a segment (20°) of an annulus having bevelled edges. The excitation time in minutes was, therefore, 20/360 × speed of the disc (coated with phosphor) in revolutions per minute. The disc is driven through a gear train of 2 : 1 ratio by a specially adapted generator (12 volts d.c. input, 100 volts 1 amp. output). The speed of the generator could be varied by means of a resistance (1–10 ohms) in the driving commutator circuit; the loading of the secondary was provided by a low-resistance voltmeter whose reading was proportional to the speed of the motor when the field current was maintained constant. The absolute speed was determined both by a stroboscopic method and by direct counting, and it was found that uniform speeds of rotation were obtained. The glass disc was coated by spraying on to a thin adhesive layer; the combined thickness of the adhesive and phospher layers (measured by an interference method) is *c.* 0·08 mm. and the particle size distribution of the phosphor is

Diam. (μ)	8	8–16	16–23	23–32	32–45	45–64
No. (%)	60·4	20·6	10·6	5·2	2·3	0·9

To enable observations to be made at temperatures up to 100° c., a thermally-insulated and light-tight jacket was fitted over the rotating disc. Into a circular aperture directly opposite the disc was fitted the lens of a quarter-plate field camera used at double extension. All joints were lined with light-proof black felt so that no extraneous light could enter the camera. Filters could be incorporated behind the lens in the camera; throughout this series of experiments a sodium nitrite solution was used to remove any λ3650 radiation which had entered the camera and a Wratten No. 39 filter to remove a faint green band occurring in the luminescent emission. Half of the disc could be photographed; each trace was produced by many (of the order of 100) revolutions of the disc, the first few activations being in no case recorded. Ilford H.P.3 plates were used throughout and a standard procedure adopted for their development. Using previously calibrated neutral filters, the plate characteristics were determined directly for the equilibrium fluorescence on each plate, thus eliminating the effect of slight differences in development conditions. The characteristics of H.P.3 plates at several wave-lengths in the range of the luminescent emission were also measured; using these, it was possible to allow for the effect of changes in the spectral distribution of the luminescent emission during build-up and decay. To avoid errors due to failure of the reciprocity law, the exposure time was made equal to that of the corresponding trace. By photographing a strip of standard intensity on each plate, the intensities calculated for the different plates were obtained in terms of a common scale factor. A representative positive of the type of trace obtained is illustrated in figure 1. To measure the transmission of the circular trace, a Hilger microphotometer was adapted for use with a rotating stage. Measurements could then be made at intervals of 30 minutes of arc, corresponding to a time interval of 0·00139 sec. for a speed of 60 revolutions per minute. The usual precautions were taken to ensure constant intensity from the microphotometer lamp; the relationship between light intensity and response of the photocell galvanometer combination was linear. Density determinations were made at a number of intervals around the trace and converted to luminescent intensity values.

The method outlined here has several advantages over those employed previously. Most previous investigators (see, for example, Beese (1939)) employed direct measurement on a rotating disc or cylinder coated with phosphor. The slit-widths used with the photocells covered a considerable time range, a factor which automatically introduces a large error, particularly at the shorter times where the intensity is changing rapidly. The electron multiplier-cathode ray oscillograph method employed by de Groot (1939a) and others is only suitable for accurate measurements over a period of 20 milliseconds or so.

Furthermore, the method outlined here has the advantage of furnishing a permanent record of the phosphorescence.

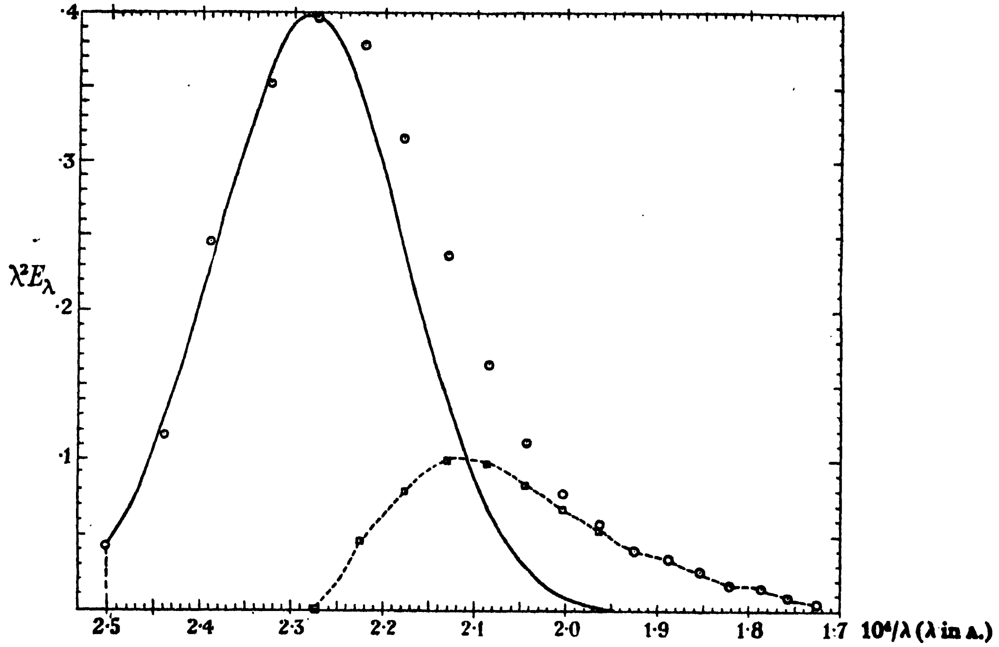


Figure 2. Spectral energy distribution of silver-activated zinc sulphide. ⊙ Experimental points.

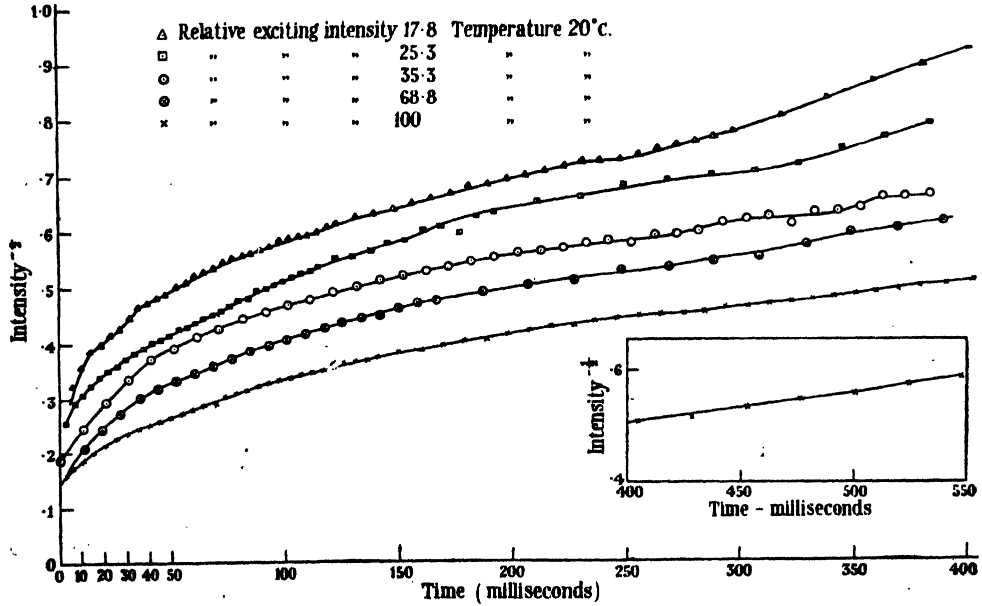


Figure 3. Plots of $y^{-\frac{1}{2}}vt$ for various exciting intensities and temperatures.
(*a*) Temperature 20° c.

§ 3. RESULTS

The crystal phosphor used in these investigations was a zinc sulphide, activated by *c.* $7 \cdot 5 \times 10^{-3}$ % silver, but containing also some ($\leqslant 1 \times 10^{-4}$ %) copper impurity, which caused further independent activation, whilst small quantities of iron ($\leqslant 15\times 10^{-5}$ %), nickel ($\leqslant 2\times 10^{-5}$ %) and chlorine ($\leqslant 150\times 10^{-5}$ %) were probably also present. The spectral energy distribution of the light emitted by the phosphor when excited by almost monochromatic radiation of wave-length 3650 A. is given in figure 2. The significance of the overlapping continua present in this distribution will be discussed later.

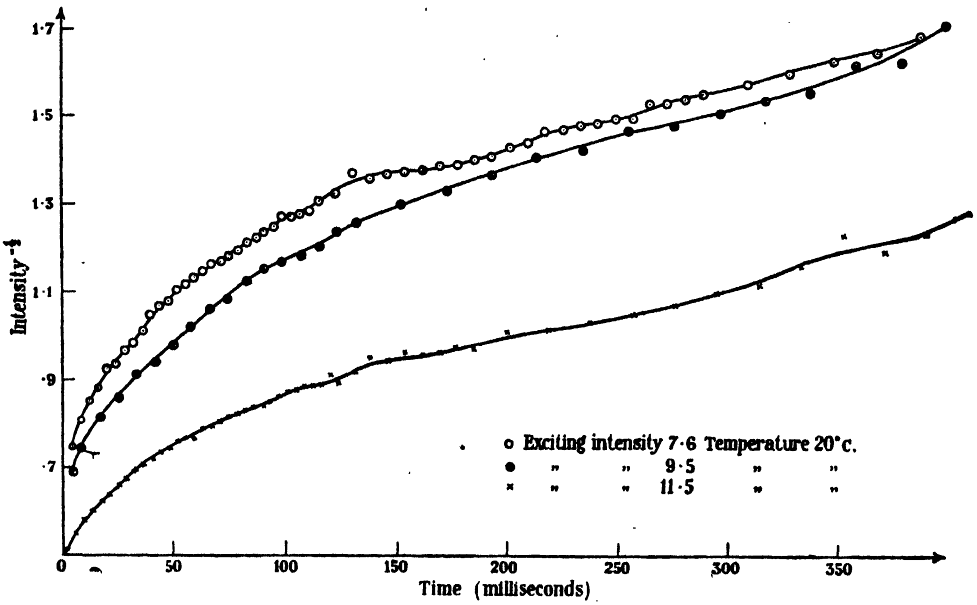


Figure 3 (*b*). Temperature 20° c.

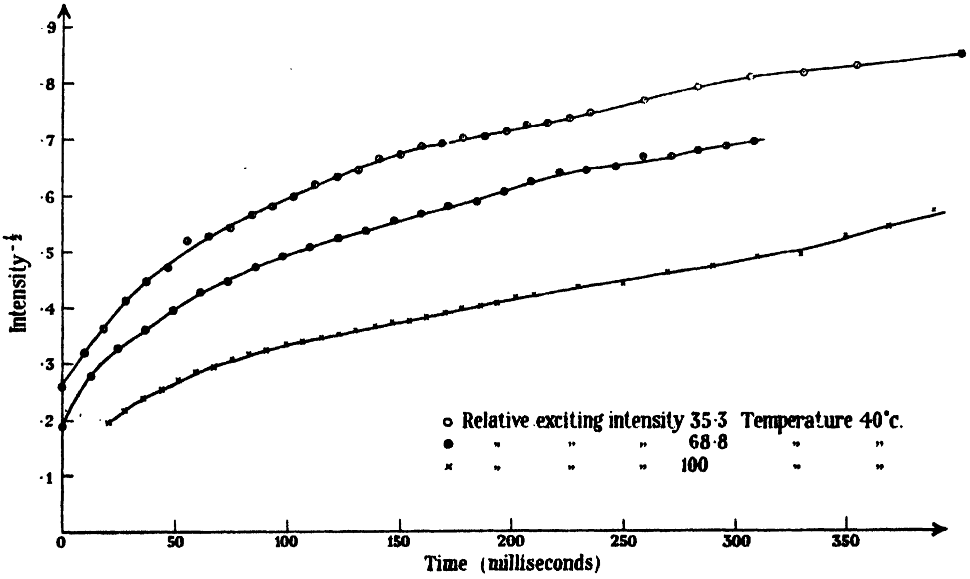


Figure 3 (*c*). Temperature 40° c.

The blue luminescent intensity-time relationships were measured over a range of exciting intensities at temperatures from 20° c. to 80° c. At some of the higher temperatures there were periods in which the density of blackening

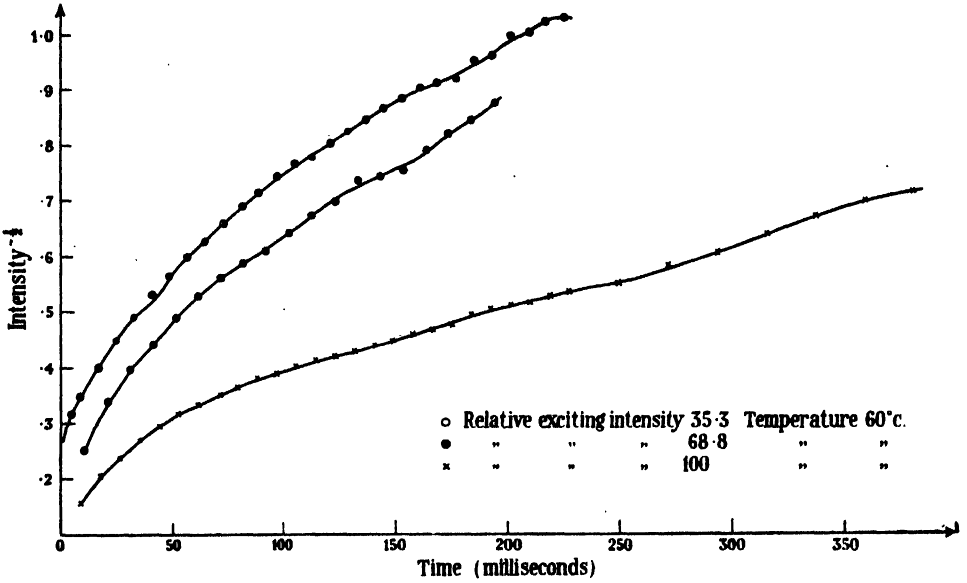


Figure 3 (*d*). Temperature 60° c.

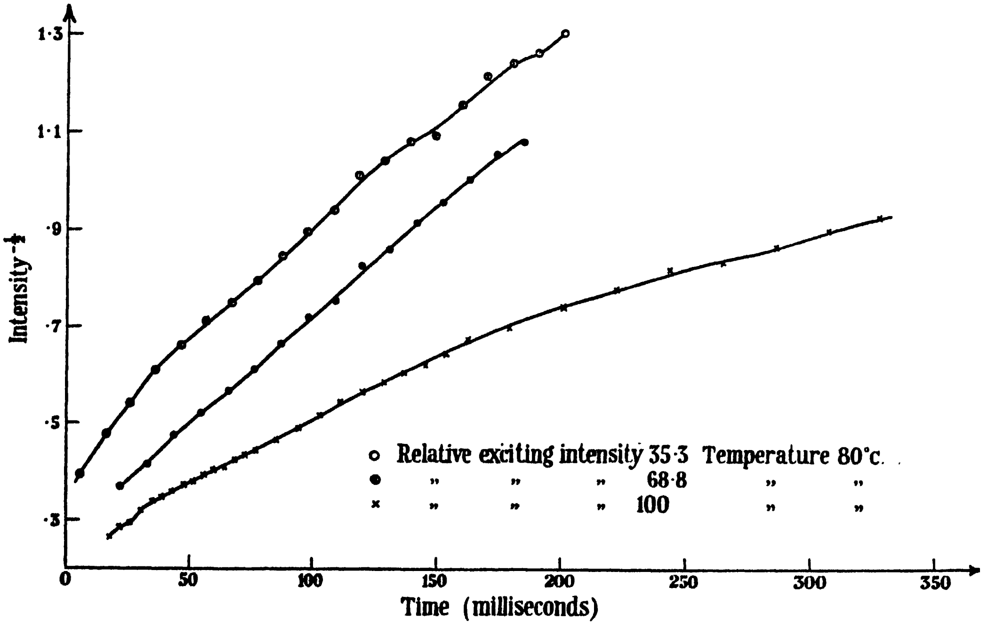


Figure 3 (*e*). Temperature 80° c.

of the photographic plate was too great to measure, so that the corresponding observations are missing. Owing to the smallness of the density of blackening at the longer times, the corresponding intensities may be unreliable. To diminish the random errors, all the observations except some extreme ones were smoothed by the least squares formula, which uses seven consecutive results to smooth the middle one of the seven. In order to test the theory of § 5, the $-\frac{1}{2}$ powers of the phosphorescent intensity are plotted against time in figure 3. The curves are drawn through the smoothed values and all the unsmoothed points are also plotted. The range of the observations, the times of activation and the corresponding values at 20° c. of the integral of the luminescent intensity with respect to time from the beginning of the build-up to infinity are given in table 1; they are discussed in § 5.

§ 4. PHYSICAL THEORY

There has been considerable argument concerning the function of the activating impurity in crystal phosphors. It seems fairly certain, however, that the activator atoms or ions, as the case may be, occupy interstitial positions in the crystal and that absorption of radiation of wave-length greater than 3340 A. in zinc sulphides occurs at intestitial atoms, whether they be excess zinc atoms or those of the added activating impurity (Seitz, 1938; 1939 a and b). The suggestion that an exciton is formed by absorption by a negative lattice ion (Kitchener, 1939; Riehl, 1939) is at variance with the experimental facts. Milner (1939) has shown that absorption by lattice ions would lead to a band width much greater than 1 ev. Furthermore, the quantum efficiency for irradiation with λ 3650 is *c.* 1, a value hardly to be expected if absorption takes place at lattice ions some distance from the source of fluorescent emission, namely the interstitial atoms. Moreover, evidence that electrons are excited to the conduction band is provided by the work of Gudden and Pohl (1921) : they found that photoconductivity accompanies absorption by λ 3650. We therefore consider, in agreement with Seitz (1938, 1939 a and b), that the energy level occupied by an electron in an activator atom lies in the region between the highest filled band and the conduction band of the crystal at a point determined by the type of activator atom, and that excitation by absorption of λ 3650 causes an electronic transition from this level to the conduction band. There is agreement that the positive hole left behind by an electron generally remains localized at the activating centre and that the emission process occurs at the activating centres. There is some divergence of opinion as to whether the spectral distribution of the emission is a function of the activating centre, or of the matrix, or of both. However, silver-activated zinc sulphides often have a band maximum at 4400 A. and copper-activated ones, a green emission. Both Ag- and Cu-activated ZnS sometimes show a band maximum at 4650 A. (Leverenz and Seitz (1939) and Rees (1942)); the emission spectra of pure zinc-sulphide phosphors also have maxima at this wave-length, and their emission is attributed to interstitial zinc (Seitz, 1939 a). We have therefore endeavoured to identify the centres in "our" phosphor by analysing the relationship between the energy ($E_\lambda d\lambda$) emitted in the wave-length range $\lambda \rightarrow \lambda + d\lambda$ and the wave-number ($1/\lambda$). This relationship (see figure 2) has a maximum at *c.* 4400 A. and a long wave-length tail extending to *c.* 5800 A. Over the range 4000–4400 A., $\lambda^2 E_\lambda$ is Gaussian, and hence, following Henderson (1939), we assume that this part is due to a single centre which we identify as silver. It can be seen from figure 2 that subtraction

Table 1. Times of activation t_0, values of $\int X\,dt$, and estimated values of $\dfrac{\int X\,dt}{I t_0}$ over build-up and decay

(I denotes exciting intensity; t denotes time of last observation)

		$\int X\,dt$. Temperature 20° c.							Temperatures (°c.)					
									40°		60°		80°	
I	t_0	Over build-up	Over observed decay	Remainder (extrapolated from $1/b_1(b_1t+C_1)$	Total	$\dfrac{\int^{\infty} X\,dt}{I t_0}$	t	I	t_0	t	t_0	t	t_0	t
100	96·0	3498·5	4000	3136	10635	1·108	548	100	79·4	449	87·6	381	85·4	329
68	70·8	1956	2535	2843	7334	1·524	452	68	122·4	308	102·0	194	108·8	184
35·3	70·7	1164·5	1750	1829	4744	1·902	384	35·3	94·4	401	80·4	225	103·0	201
25·3	79·0	750·8	1425	1353	3529	1·579	384							
17·8	78·0	602	1062	1123	2787	2·008	464							
11·0	78·0	200·3	488	643	1331	1·552	403							
9·54	74·0	111·1	270	365	746	1·057	462							
7·20	79·0	91·2	234	364	689	1·211	388							

N.B.—Unit of time is 1 millisecond.

of the silver band * from the complete distribution leaves an asymmetrical band in which *c.* 20% of the total energy is emitted. This complex consists mainly of a single band with a maximum at *c.* 4680 A., and there is also a broad band with a maximum in the green. We therefore consider the phosphor to contain three types of centre: Ag, Zn and Cu in order of increasing wave-length of their maxima.

There is a divergence of opinion about the history of the electron between the time when it is excited into the conduction band and the time when it returns to an activating centre. Many workers (e.g. Lewschin and Antonow-Romanowsky (1934)) consider that the time is spent in wandering freely through the conduction band, and have shown that such behaviour leads to a bimolecular type decay. Randall and Wilkins (1945 b) and Garlick and Wilkins (1945) have, however, pointed out that for decays of the duration commonly found, this theory would lead to rather small values of the cross-section of capture of an electron by a luminescence centre; also, that the decay rate would be independent of the time of activation, and would increase slowly with temperature, and that there would be an inflexion in the build-up curve. They obtained curves of this type for some phosphors over a limited temperature range, but most of their observations do not support the simple biomolecular theory. They suggest that the electron spends the greater part of its lifetime in the excited state in a trap (or traps) from which it can be released by absorbing thermal energy from the lattice: the number and depth of the traps are considered to be a function of the structure of the phospher only.

However, we have found that, assuming two types of activating centre, a bimolecular law can be fitted to each of our decay curves with the exception, in some cases, of a few points at the beginning and end. The variation with temperature and exciting intensity of the four independent constants obtained from each curve cannot be explained on either the simple bimolecular theory or on a theory incorporating the ideas of Garlick and Wilkins (1945) on bimolecular type decays. Nearly all our results are, however, explained by introducing a new trapping mechanism.

§ 5. NUMERICAL ANALYSIS OF THE DECAY RESULTS ASSUMING THAT THERE ARE TWO TYPES OF ACTIVATING CENTRE (SILVER AND ZINC)

We shall denote by n_1 and n_2 the number of excited activating centres per unit volume of the first and second kinds.

The excited electrons wander freely through the conduction band. Luminescence is emitted when such an electron is captured by an ionized activating centre and falls to a lower energy level. The probability that a particular activating centre will be filled in unit time is proportional to the concentration of the electrons (of number n) per unit volume in the conduction

---

* Henderson (1939) has found that one kind of centre gives rise to a spectral distribution whose E_λ values (with the exception of the smallest ones) are well represented by the Gaussian expression

$$E\lambda = \frac{A'}{\lambda^2\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(\frac{10^4}{\lambda} - \frac{10^4}{\lambda_0}\right)^2}.$$

The experimental observations between 4000–4400 A. lead to $\sigma = 0{\cdot}1046$, $\lambda_0 = 4386$ A. ($A' = 1$) for the silver band of our phosphor.

band round the centre, and is assumed to be $\beta_1 n$ for an activating centre of the first kind and $\beta_2 n$ for an activating centre of the second kind.

Then for a unit volume of the lattice, within which we can neglect the effect of variations of n,[*] the equations for the decay are

$$dn_1/dt = -\beta_1 n_1 n, \qquad \ldots\ldots(5.1)$$

$$dn_2/dt = -\beta_2 n_2 n, \qquad \ldots\ldots(5.2)$$

$$n = n_1 + n_2. \qquad \ldots\ldots(5.3)$$

To apply these equations to the analysis of our observations, it is necessary to relate the observed luminescent intensity to the rate ($-dn_1/dt - dn_2/dt = X_1 + X_2 = X$, say) at which quanta are emitted by the two kinds of centre. If $(E_{1t\lambda}/h\nu)d\lambda$, $(E_{2t\lambda}/h\nu)d\lambda$ quanta per second per unit volume are emitted by centres of the first and second kind respectively at time t in the wave-length range $\lambda \rightarrow \lambda + d\lambda$, the recorded luminescent intensity (y) is $A'\int F_\lambda(E_{1t\lambda} + E_{2t\lambda})d\lambda$. A' is the same for all observations, and is the ratio of the recorded intensity of the standard strip to the value of the integral for this strip. F_λ is a ratio[†] depending on the transmission factor of the blue filter and the deviation of the characteristic corresponding to the spectral distribution for the standard strip from the characteristic at wave-length λ. The band spectra of the centres are unlikely to change shape during the time variation of the luminescence, so we can put

$$E_{1t\lambda} = \frac{X_1 E_{1\lambda}}{\int \frac{E_{1\lambda}}{h\nu} d\lambda}, \qquad E_{2t\lambda} = \frac{X_2 E_{2\lambda}}{\int \frac{E_{2\lambda}}{h\nu} d\lambda},$$

where $E_{1\lambda}$ and $E_{2\lambda}$ are independent of time and proportional to the heights at wave-length λ of the repsective blue bands to which they correspond. Then

$$y = A'\left[X_1 \frac{\int F_\lambda E_{1\lambda} d\lambda}{\int \frac{E_{1\lambda}}{h\nu} d\lambda} + X_2 \frac{\int F_\lambda E_{2\lambda} d\lambda}{\int \frac{E_{2\lambda}}{h\nu} d\lambda}\right]. \qquad \ldots\ldots(5.4)$$

This can be written as

$$Ay = -dn_1/dt - A_F dn_2/dt. \qquad \ldots\ldots(5.4')$$

A_F is found to be 2 if centres of the second kind are silver; it will, of course, be $\frac{1}{2}$ if centres of the second kind are zinc.

The equations (5.1)→(5.4′) can be solved completely. In discussing them it is convenient to introduce

$$n_{10} = \text{value of } n_1 \text{ at } t = 0;\ n_{20} = \text{value of } n_2 \text{ at } t = 0;\ \zeta = n_{20}/n_{10};$$
$$\rho = \beta_2/\beta_1;\ u = n_1/n_{10}.$$

---

[*] The effect of variations of n within the layer is considered in Appendix B.

[†] It is the product of the transmission factor of the filter and a response factor obtained from H.P.3 characteristics measured at various wave-lengths (§ 2); the response factor appears to be independent of intensity over a wide range of intensities. The relative values of F_λ are 38, 38, 43, 35, 9·2, 1·53 and 0 at $\lambda = 4000$, 4200, 4400, 4600, 4800, 5000 and 5200 A. respectively.

We can immediately obtain n_2 in terms of n_1 by dividing (5.2) by (5.1):

$$n_2 = n_{20} u^{\rho}. \qquad \ldots\ldots(5.5)$$

Hence X and t can be obtained in terms of n_1 only:

$$X = \beta_1 n_{10}^2 u^2 (1 + \zeta u^{\rho-1})(1 + \rho\zeta u^{\rho-1}) \qquad \ldots\ldots(5.6)$$

and

$$Ay = \beta_1 n_{10}^2 u^2 (1 + \zeta u^{\rho-1})(1 + A_F \rho\zeta u^{\rho-1}), \qquad \ldots\ldots(5.6')$$

$$t = \frac{1}{\beta_1 n_{10}} \int_u^1 \frac{du}{u^2 + \zeta u^{(1+\rho)}}. \qquad \ldots\ldots(5.7)$$

We shall designate the centres so that $\rho > 1$. Then n_2/n_1 ultimately becomes negligible. If $\rho > 2$ it is best to integrate (5.7) by parts and write it as

$$\begin{aligned}\beta_1 n_{10} t &= \frac{1}{u(1 + \zeta u^{\rho-1})} - \frac{1}{1+\zeta} - \zeta(\rho - 1)\left[\int_0^1 - \int_0^u \frac{u^{\rho-8}\,du}{(1 + \zeta u^{\rho-1})^2}\right] \\ &= \frac{1}{u(1 + \zeta u^{\rho-1})} - C + \zeta(\rho - 1) S_u, \qquad \ldots\ldots(5.8)\end{aligned}$$

where

$$S_u = \int_0^u \frac{u^{\rho-3}\,du}{(1 + \zeta u^{\rho-1})^2} \qquad \ldots\ldots(5.9)$$

and

$$C = \frac{1}{1+\zeta} + \zeta(\rho - 1) S_1. \qquad \ldots\ldots(5.10)$$

Thus S_u is a power series with terms $u^{\rho-2}$, $u^{2\rho-3}$, etc. So after a long time

$$\beta_1 n_{10} t + C = 1/u \qquad \ldots\ldots(5.11)$$

and

$$y^{-\frac{1}{2}} = b_1 t + C_1, \qquad \ldots\ldots(5.12)$$

where

$$b_1 = \sqrt{(\beta_1 A)} \qquad \ldots\ldots(5.13)$$

$$C_1 = \frac{C b_1}{\beta_1 n_{10}}. \qquad \ldots\ldots(5.14)$$

The linear relationship (5.12) is satisfied by the observations over a wide range of values of t (given in table 2 *a*) in all decay curves. Sometimes the last few values are too large at 20° c. and too small at 60° c.; these in any case are experimentally unreliable. b_1 and C_1 were thus obtained numerically (table 2 *a*).

The next stage was to test (5.6) and (5.7) at shorter times by including the $u^{\rho-1}$ terms. We can expand (5.8) in powers of $\zeta u^{\rho-1}$, and when this is small, but not negligible, there will be observations for which

$$y(b_1 t + C_1)^2 - 1 \doteqdot B(b_1 t + C_1)^{1-\rho}, \qquad \ldots\ldots(5.15)$$

where

$$B = \zeta\left(1 + \rho A_F + \frac{2}{\rho - 2}\right)\left(\frac{C_1}{C}\right)^{\rho-1} \qquad \ldots\ldots(5.16)$$

Log $[y(b_1 t + C_1)^2 - 1]$ was therefore plotted against $\log(b_1 t + C_1)$ for all observations (except the 80° c. ones, for which $y^{-\frac{1}{2}}$ did not deviate sufficiently from $b_1 t + C_1$). A specimen curve is given in figure 4. It was then clear that at 20°, 40° and 60° c. (5.15) was correct from *c.* 20 milliseconds onwards. ρ (table 2 *a*) was obtained from the slope of the lines; its values are probably not so accurate as those of b_1 and C_1 since the values of $y(b_1 t + C_1)^2 - 1$ are small differences. However, within these limits ρ has the same values for different exciting intensities, viz. 20 at 20° c., 17·5 at 40° c., and 11·5 at 60° c. This is new evidence for the two-centre theory.

The intercepts of the figure 4 type curves give $\log B$ numerically (table 2 *a*).

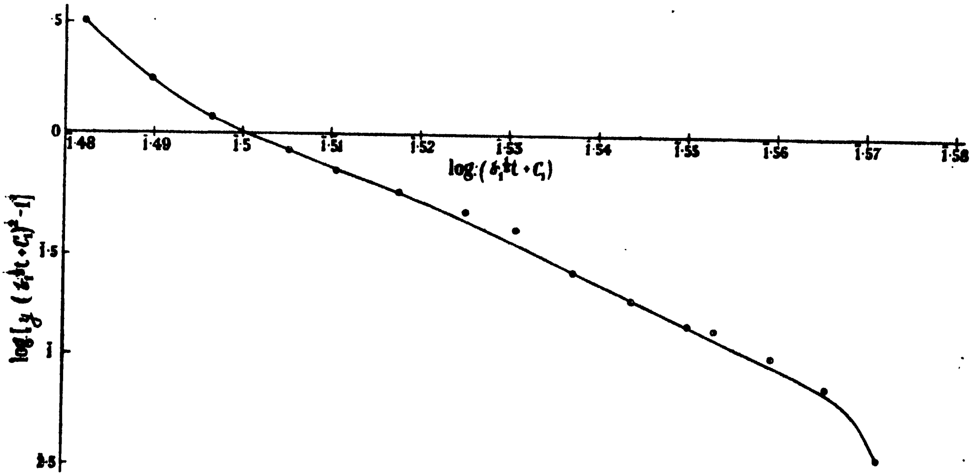


Figure 4. A representative plot of $\log\{y(b_1t+C_1)^2-1\}$ against $\log(b_1t+C_1)$.

Table 2 *a*. Constants obtained directly from the decay observations

Temperature (°c.)	I	b_1	C_1	Times between which $y^{-\frac{1}{2}}=b_1t+C_1$ (seconds)	y_0*	$\log_{10}B^{\delta}$	ρ
20	100	0·535	0·303	0·14–0·55	51·8	$\bar{1}$·505	19·9
	68·8	0·553	0·386	0·16–0·39	49·0	$\bar{1}$·605	19·8
	35·3	0·780	0·402	0·13–0·35	28·5	$\bar{1}$·616	20·6
	25·3	0·925	0·444	0·12–0·39	16·7	$\bar{1}$·672	19·3
	17·8	0·950	0·496	0·11–0·32	11·9	$\bar{1}$·697	20·4
	11·0	1·24	0·754	0·11–0·40	4·25	$\bar{1}$·873	20·1
	9·5	1·54	1·067	0·14–0·40	2·35	0·033	19·8
	7·6	1·58	1·124	0·10–0·28	2·00	0·047	21·8
40	100	0·780	0·256	0·09–0·34	—	$\bar{1}$·4405	17·7
	68·8	0·880	0·436	0·19–0·40	28·5	$\bar{1}$·6585	17·3
	35·3	0·915	0·538	0·14–0·31	15·0	$\bar{1}$·7460	17·5
60	100	0·980	0·308	0·15–0·30	—	$\bar{1}$·514	12·1
	68·8	2·49	0·388	0·08–0·20	21·0	$\bar{1}$·6255	11·4
	35·3	2·54	0·496	0·09–0·18	15·0	$\bar{1}$·720	11·0
80	100	2·58	0·246	0·03–0·16	—	—	—
	68·8	4·58	0·265	0·02–0·18	—	—	—
	55·3	4·72	0·434	0·03–0·14	—	—	

* y_0 is the value of y at $t=0$.

(5.16) can now be solved for ζ. This is easy if more manageable expressions for C are first found: they are obtained by substituting expansions for S_u in powers of $\delta \equiv 1/(\rho - 1)$ (Appendix A) in (5.8), viz.

$$\beta_1 n_{10} t + C = \frac{1}{u}[1 + \delta \log_e(1 + \zeta u^{\rho-1}) + 0(\delta^2 \zeta u^{\rho-1})] \text{ if } \zeta u^{\rho-1} < 1 \qquad \ldots\ldots(5.17)$$

$$= \frac{\pi\delta\zeta^{\delta}}{\sin \pi\delta} + \frac{1}{u}\left[\delta \log\left(1 + \frac{u^{1-}}{\zeta}\right) - 0\left(\delta^2 \frac{u^{1-\rho}}{\zeta}\right)\right] \text{ if } \zeta u^{\rho-1} > 1. \qquad \ldots\ldots(5.18)$$

Then putting $t = 0$ (i.e. $u = 1$) immediately yields C.* It is easily seen that the ζ values corresponding to $A_F = \frac{1}{2}$ will differ from those corresponding to $A_F = 2$. It is therefore necessary to decide which value to take for A_F. The ratio of the contribution to the luminescence of centres of the second kind to that of the first kind at $t = 0$ is $\rho\zeta$, which (from the approximate values of the ζ obtained by putting $C = 0$) tends to increase with exciting intensity for both values of A_F and, for $I = 100$, $T = 20$°c., $= 1{\cdot}4$ if $A_F = 2$ and $7{\cdot}0$ if $A^F = \frac{1}{2}$. Now the greater part of the fluorescent emission (figure 2) measured at intensities somewhat higher than those of our observations is due to silver centres, and so it seems reasonable to identify centres of the second kind with silver and those of the first kind with zinc,† i.e. $A_F = 2$.

The ζ being known, (5.17) with (5.6) also gives values of X and t over all the decay period. It was used in particular to calculate the times (corresponding to $u^{\rho} = \frac{1}{2}$) taken for the number of positive holes of the centre with the faster decay (silver) to fall to half its initial value. The first term suffices for calculating the much longer half-value times (corresponding to $u = \frac{1}{2}$) for the zinc centres. These times are tabulated with other derived constants of physical interest, including the ratio of the observed to calculated values of y_0, in table 2 *b*.

The constants (tables 2*a* and 2*b*) vary with I in an unexpectedly complicated way. At 20° c., $\beta_1 n_{10}$ appears to vary periodically with I: its values at 40, 60° and 80° c. are not inconsistent with a similar behaviour. Moreover, almost all other constants as functions of I appear to have periodic terms superimposed on a simpler function. If these periodicity effects are real, they might be explained if the differential equations governing the build-up have periodic terms: in Appendix D we show that this may be so under certain conditions.

There are, however, several features of the results which cannot be explained. Thus the ratio of the estimates by two different methods of quantities proportional to the total number of excited electrons from blue centres are different for each set of observations at 20° c. The first method of estimation is given in the build-up section, where it is shown that the total number of excited electrons is proportional to the product of the exciting intensity (I) and the time of activation (t_0). The second method depends on the fact that the total number of electrons excited from blue centres is $\int X\,dt$ integrated over build-up and decay. The value of this integral over the decay could be obtained by

---

* It is interesting to note that when δ is small, $C \simeq 1$ even if ζ is large. This means that the final decay equation is changed only slightly by the presence of a second centre with a considerably faster β even if the number of its holes is large.

† The more so as rough measurements of the spectral distribution indicate that the ratio of the contribution from zinc and silver centres decreases with exciting intensity.

Table 2 *b*. Derived constants

Temperature (°c.)	I	ζ	C	$\beta_1 A$	$\beta_1 n_{10}$	$\frac{\beta_1 n_{10}^2}{A}$	$\frac{n_{10}}{A}$	$\frac{n_{20}}{A}$	$\frac{\text{Observed } y_0}{\text{Calculated } y_0}$	Times (sec.) $u=\frac{1}{2}$	Times (sec.) $\frac{n_2}{n_{20}}=\frac{1}{2}$*
20 $\rho=20$	100·0	0·072	1·0039	0·286	1·773	10·99	6·20	0·447	1·134	0·5618	0·0198
	68·8	0·058	32	0·306	1·437	6·76	4·70	0·273	2·063	0·6937	0·0244
	35·3	0·043	24	0·608	1·945	6·22	3·20	0·137	1·616	0·5123	0·0180
	25·3	0·077	40	0·856	2·084	5·08	2·44	0·189	0·748	0·4779	0·0169
	17·8	0·027	15	0·902	1·919	4·08	2·13	0·0574	1·366	0·5243	0·0183
	11·0	0·021	12	1·538	1·647	1·765	1·072	0·0225	1·282	0·6665	0·0213
	9·5	0·031	17	2·372	1·445	0·881	0·610	0·0189	1·155	0·6908	0·0244
	7·6	0·033	18	2·496	1·408	0·794	0·564	0·0186	1·052	0·7089	0·0250
40 $\rho=17\cdot5$	100·0	0·105	1·0065	0·608	3·067	15·47	5·04	0·529	—	0·3240	0·0121
	68·8	0·061	38	0·774	2·027	5·31	2·62	0·160	1·614	0·4914	0·0190
	35·3	0·082	50	0·837	1·710	3·49	2·04	0·167	1·027	0·5818	0·0222
60 $\rho=11\cdot5$	100·0	0·084	1·0084	0·960	3·209	10·73	3·34	0·281	—	0·3090	0·0181
	68·8	0·113	113	6·20	6·490	6·80	1·048	0·118	0·771	0·1524	0·0087
	35·3	0·082	82	6·45	5·168	4·14	0·801	0·0657	1·161	0·1918	0·0113
80	100·0	—	—	6·80	10·5	16·55	1·58	—	—	0·0952	—
	68·8	—	—	21·0	17·3	14·24	0·826	—	—	0·0378	—
	35·3	—	—	23·3	10·9	5·32	0·489	—	—	0·0918	—

* $\frac{n_2}{n_{20}}=\frac{1}{2}$ corresponds to $u=0\cdot9658$ at 20° c., 0·9612 at 40° c. and 0·9415 at 60° c.

extrapolation from the last recorded observations (table 1); its values over the build-up could only be found for the 20° C. observations as the build-up results at the other temperatures were not complete. The values of the ratio (table 1) indicate that the decay must be slower than a -2 power law at long times. It is significant that measurements of the phosphorescent intensity of ZnS at long times published by Johnson (1939), Fonda (1945), Randall and Wilkins (1945 c) and Jesty (1946) do not satisfy the law of (5.11). Another unexplained feature is that the observed y_0 values do not agree with the calculated ones. Again, as Garlick and Wilkins (1945) have pointed out, the β should be proportional to the square root of the absolute temperature, but they vary with temperature more rapidly than that. Finally, the most serious difficulty is that the β should not depend on I, but they are, in fact, roughly proportional to $1/I$.

In the next section we shall show how these difficulties may be accounted for and the constants reinterpreted by allowing for the presence of trapping centres.

Two other factors have been neglected in developing the theory. Thus the phosphor layer is thick enough to cause considerable variations in the conditions at different depths. However, when β_1 decreases as n_{10} increases, so that $\beta_1 n_{10}$ does not vary greatly either with I or with depth in the layer, the law still holds. It is valid in particular for the amount of variation of $\beta_1 n_{10}$ at 20° C. in table 2 *b*. These statements are amplified in Appendix B. The effect of the copper centres has also been neglected. If the number of activated green centres per unit volume is n_3, we should put

$$\frac{dn_3}{dt}=\beta_3 n n_3, \qquad \ldots\ldots(5.19)$$

$$n=n_1+n_2+n_3. \qquad \ldots\ldots(5.20)$$

Equations (5.1) and (5.2) are unchanged except that n is given by (5.20) and (5.4) is unaltered since the green luminescence is filtered out. A justification for neglecting n_3 is suggested in the next section.

§ 6. PROPOSED THEORY FOR PHOSPHORS HAVING MORE THAN ONE TYPE OF ACTIVATING CENTRE AND TRAPPING CENTRES

The phenomenon of thermoluminescence is well known. There is general agreement that it is due to electrons being trapped after excitation in some regions of the crystal (trapping or metastable levels) from which they can only be released by receiving thermal energy from vibrations of the lattice ions. Randall and Wilkins (1945 a, b) have studied the subject extensively: for example, they excited a phosphor for some time at low temperature and measured the phosphorescent emission when it was heated at a constant rate; they consider that the intensity-temperature relationship (glow curve) so obtained approximately corresponds to the trap depth distribution of the phosphor. The physical picture of the trapping levels is not very satisfactory; they have been associated with vacant negative ion lattice points, cracks and regions of strain in the crystal and surface levels.

The simplest assumption is that of de Groot (1939) and Blokhinzev (1937), who suggested that trapped electrons are released by absorbing thermal energy at a rate γl proportional to the number of electrons in traps, and that traps capture electrons at a rate $\alpha n(L-l)$ proportional to the concentration n of free electrons and the concentration of empty traps $L-l$, where the number of positive holes per unit volume is, of course, $n+l$. It is shown in Appendix C (iv) that for the two-centre case these assumptions lead at long times to the relations (5.1) and (5.2) but with the β_1 multipled by $\gamma/(\alpha L+\gamma)$. This ratio is the limiting value (R) of the ratio (r) of free electrons to holes.

Now γ is of the form $se^{-E/kT}$, where E is the trap depth, T the absolute temperature and k is Boltzmann's constant, and according to Garlick and Wilkins (1945) β_1 and β_2 increase as the square root of T. Klasens and Wise (1946) have pointed out that if T is high or the traps are very shallow, $\gamma/(\alpha L+\gamma)\simeq 1$ and $\beta_1 R$ and $\beta_2 R$ increase only slowly with temperature, while if $\alpha L \gg \gamma$, electrons when excited spend most of the time in traps and $\beta_1 R$ and $\beta_2 R$ will increase more rapidly with temperature. However, there seems no reason why the ratio of the βR should change with temperature. Also it seems quite impossible for $\beta_1 R$ and $\beta_2 R$ to change with I.

We shall now show that these difficulties may be resolved by postulating another trapping mechanism. Since a change in I produces a change in the crystal only at the activating centres, we believe that there must be traps closely associated with the excited centres. We suggest, therefore, that when a positive hole is formed on an activating centre, shallow secondary potential holes which can trap electrons are produced in its neighbourhood: these secondary holes disappear when the primary positive holes are filled. Electrons can be released from the secondary holes by absorbing vibrational energy from neighbouring ions. An electron so released may either fall into the primary hole or wander through the lattice. The process is somewhat analogous to that of an electron digging its own hole (Landau (1933), Hippel (1936), Gurney and Mott (1937) and Mott (1937)).

It was assumed that there were $\lambda_j n_j (j=1, 2, 3)$ traps per unit volume associated with positive holes of the jth kind and that $\gamma_j l_j$ electrons vibrated out of them per second, of which $\gamma_j l_j p_j$ went straight into positive holes and $\gamma_j l_j(1-p_j)$ into the conduction band. It seemed likely that the rate at which free electrons fell into positive holes would depend on whether or not the associated traps were empty. We therefore considered in some detail the behaviour of n_j and l_j as functions of time when there was one trap per hole and the hole could not be filled by a free electron when the associated trap was occupied (Appendix C (i)). Of the various types of solution there was one which gave bimolecular laws at long times of the form of (5.1) and (5.2) but with β_j replaced by $\beta_j+\alpha_j p_j$. The proportion of excited electrons in traps could then be small even though free electrons were more often captured by traps than by holes. In the more general case, without the above restrictions, it was still found that β_j was replaced by $\beta_j+\lambda_j \alpha_j p_j$*: the result is independent of γ_j if γ_j is large and the traps are shallow, and physically this is equivalent to saying that traps delay the recombination.

It is known from observation that the decay of the green phosphorescence is much slower than that of the blue. Hence β_3 must be small compared with β_1 and, therefore, from the above theory either α_3 or p_3 must be small compared

---

* Cf. equations (C.24) and (E.3) on pp. 497 and 500 respectively.

with α_1 and p_1 respectively. If p_3 is small, the traps associated with copper centres will behave like shallow de Groot-type traps with respect to the blue luminescence during the period of our observations. So their effect is to multiply the β_3 for the other centres by $\gamma_3/(\lambda_3\alpha_3n_{30}+\gamma_3)$, where n_{30} is the value of n_3 at short times regarded as constant. Now n_{30} will be approximately proportional to I, and hence the observed increase of $1/\beta_1$ and $1/\beta_2$ with I is immediately explained.

The result of the new trapping theory is, therefore, that the quantities tabulated as β_1A and β_2A in table 2 *b* should be tabulated as $\beta_1'AR$, $\beta_2'AR$ or

$$(\beta_1+\lambda_1\alpha_1p_1)A\bigg/\left(1+\frac{\lambda_3\alpha_3n_{30}}{\gamma_3}\right),\quad (\beta_2+\lambda_2\alpha_2p_2)A\bigg/\left(1+\frac{\lambda_3\alpha_3n_{30}}{\gamma_3}\right)\ \text{respectively.}$$

The ratio ρ is now β_2'/β_1' and is still independent of I. It is difficult to predict how it should change with T since the depth of the secondary potential holes as well as the energy distribution of the trapped electrons will depend on T. However, it is unlikely* that p_1 and p_2 will be invariant with T.

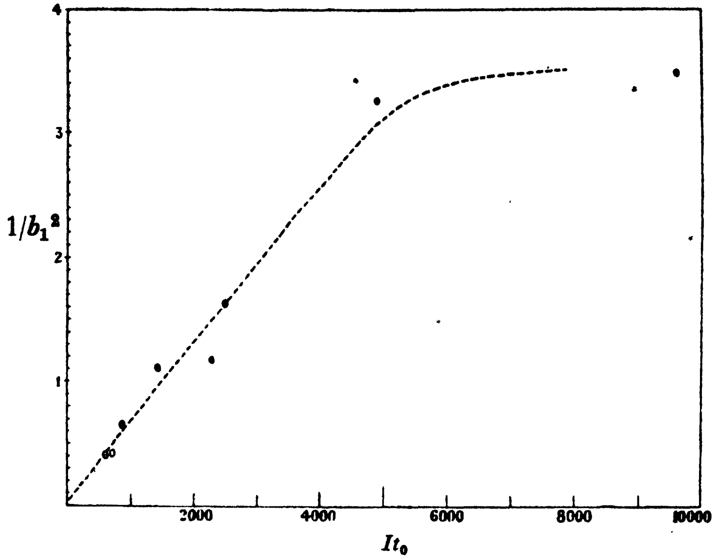


Figure 5. The variation of $1/b_1^2$ with intensity of exciting radiation.

There are enough values of $\beta_1'AR$ at 20° c. to justify examining the change with I more fully. From figure 5 it will be seen that apart from the periodic term and the point corresponding to $I100$, the values do lie about a straight line which nearly passes through the origin. This indicates that $\lambda_3\alpha_3n_{30}\gg\gamma_3$ since $1/b_1^2=[1+\lambda_3\alpha_3n_{30}/\gamma_3]/\beta_1'A$, and so most of the electrons are in traps associated with copper centres. This agrees with Garlick and Wilkins' (1945) conclusion that there are relatively few free electrons after the first few milliseconds.† It may mean that the copper traps are more numerous than the others or that their cross-section is larger or that they are deeper.

---

* See also hole migration discussion (p. 490).

† Klasens and Wise (1946) showed that with one activating centre and $\alpha=\beta$

$$r-R=(r_0-R)e^{-(\beta L+\gamma)t},$$

where $r=r_0$ at $t=0$. Using de Groot's (1939) estimate of $\beta L+\gamma$ (*c.* 100), $r-R$ is already small after a time of the order of 20 milliseconds.

Since, from the ratio of their concentrations, the number of copper centres is approximately 1/100 of the number of silver ones (§ 3), it follows that either there must be very many traps associated with each activated copper centre or that the ratio of excited to unexcited centres is much greater for copper than for the other centres. This can happen if either the absorption coefficient for copper centres is larger than for the others or if the numbers of excited centres are approaching their equilibrium values, when the ratio of excited Cu to excited Ag centres is $\mu_3\beta_2'/\mu_2\beta_3'$ (Appendix E) which can be large since $\beta_2' \gg \beta_3'$.

It is also possible that positive holes are migrating through the filled band to unactivated centres (Klasens, 1946), although the process is infrequent at low temperatures. An obvious effect would be that trapped electrons would escape to the conduction band, so that $\gamma_f(1-p_f)$ would increase more rapidly with temperature when hole migration was taking place. Another effect would be to increase the proportion of excited copper centres. This is because the probability that a hole will leave a blue centre is greater than the probability of one leaving a green centre, since the green energy levels are further from the filled band than the blue ones. The second process is small compared with the first if the cross-sections of hole capture for the different centres are of the same order, since the number of copper centres is $<1\%$ of the number of other centres. Further evidence for this is that the equilibrium blue fluorescent intensity is proportional to the exciting intensity (see figure 6, §7): Klasens has shown that hole migration from blue to green centres causes the blue intensity to increase more rapidly with exciting intensity than this. It is also possible that a net migration from silver to zinc centres (whose energy levels are nearer the filled band than the silver ones are—§4*) may account for the tendency of ζ to increase with exciting intensity.

According to Randall and Wilkins (1945 b) the long-period phosphorescence is due entirely to electrons escaping from deep traps. In our case, the number so trapped must be of the same order as the number of green centres—if it were much greater, there would be a stronger long-period blue afterglow; if it were much less, the short-period green phosphorescence would be stronger. The number of electrons escaping from the deep traps will be negligible during the period of our observations, and hence their effect is to reduce the number of free electrons from $n_1+n_2+n_3-l$ to approximately n_1+n_2-l, thus justifying our neglect of the electrons from copper centres in the expression (5.3) for the number of free electrons in § 5.

The above work is based on the assumption that the traps associated with one type of centre are all of the same depth and cross-section. However, in view of the work of Randall and Wilkins (1945 a) this does not seem likely. We have therefore endeavoured to find the effect on our theory of assuming a continuous distribution of trap depths to be associated with each kind of centre.

For the zinc and silver centres, no general results were obtained, but it seemed likely (Appendix C (iii)) that, provided the final law was a power law, $\beta_1'(=\beta_1+\lambda_1\alpha_1p_1)$ generalizes to $\beta_1+\frac{\lambda_1}{n_1}\int\alpha_E p_E n_E dE$ integrated over all trap depths, where

---

* This migration from silver, of course, rapidly becomes negligible after the end of the excitation.

$n_E dE$ is the number of holes which have $\lambda_1 n_E dE$ traps of depths in the range E to $E+dE$ associated with them. α_E, γ_E and p_E refer to traps of depths E to $E+dE$ and have their usual significance. For the copper centres, regarding the associated traps as de Groot-type ones, it is found (Appendix C (v)) that if eventually the relationship between phosphorescent intensity and time can be represented by a power law, the power is always -2, and $L\alpha/\gamma$ is replaced by $\int \frac{\alpha_E L_E}{\gamma_E} dt$ integrated over all depths; all the earlier conclusions are therefore unaffected.

It also follows that a limiting power law cannot be due to traps unassociated with excited centres when the index is not -2. This is interesting in view of Randall and Wilkins' (1945 b) results on long-period decays. They found limiting power laws other than -2 experimentally in several cases: e.g. for a ZnS with silver and copper impurities, X was proportional to $t^{-1\cdot 32}$ for $t \simeq 20$ sec. – 200 min. They showed theoretically that limiting power laws of any value < -1 are to be expected if the distribution of trap depths is exponential and it is assumed that the long-period phosphorescent intensity is proportional to the rate of excape of electrons from traps. They also found that the glow curve for phosphors with limiting power laws < -1 were approximately exponential in the region mainly responsible for the long-period phosphorescence, and that the limiting power laws calculated on the assumption that these glow curves give the trap-depth distribution agreed with the measured ones in several cases. Now this theoretical interpretation of the phosphorescence in terms of trap-depth distributions is only valid if an electron enters a hole almost immediately after it escapes from a deep trap and is never retrapped. This follows naturally from our theory if electrons go straight to copper holes from associated traps (i.e. $p_{3E} \simeq 1$ when E is large). On the other hand it seems unlikely that traps of such depth are produced merely by the ionization of Cu centres. Possibly the traps are due to localized lattice distortions associated with every copper centre whether activated or not, but can only be filled by electrons that hav just been excited from the copper centres. This would explain also why the number of electrons in deep traps is about equal to the number of excited copper centres.

§7. THE BUILD-UP OF THE LUMINESCENCE

The equations for the build-up of the luminescence are difficult to discuss even when no trapping centres are involved because the effect of no kind of centre can be neglected.* So far we have only worked on the build-up equations corresponding to C.2 and C.3 (p. 495) with one trap associated with each hole.

If we assume that the number of positive holes of the jth kind is small compared with the number of such centres, then the rate at which electrons are ejected rom centres of this type is $\frac{I\mu_j}{h\nu}\,(=k_j)$, where $\frac{I}{h\nu}$ is the number of exciting photons incident on this phosphor per second and μ_j is the probability that a particular

---

* For this reason we have not included the build-up observations. At 20° C. they showed that the period of activation was never long enough for the fluorescence to reach its equilibrium value.

photon excites any activating centre. Making the further assumption that the number of electrons falling straight back into their centres is negligible, and writing q_j for the probability that an ejected electron wanders through the crystal as a conduction electron, then $1-q_j$ must be the probability that it falls immediately into the associated trap during that period. The differential equations governing the build-up of fluorescence are then

$$\frac{dn_j}{dt}=k_jq_j-(\beta_j+\alpha_j)nn_j+\gamma_j(1-p_j)l_j, \qquad \ldots\ldots(7.1)$$

$$\frac{dl_j}{dt}=k_j(1-q_j)+\alpha_jnn_j-\gamma_jl_j, \qquad \ldots\ldots(7.2)$$

where $j=1, 2,$ or 3,

$$X=k_1+k_2-\frac{d}{dt}(n_1+n_2+l_1+l_2). \qquad \ldots\ldots\ldots(7.3)$$

The linearity of the relationship between the equilibrium fluorescent and exciting intensities (illustrated in figure 6) is consistent with these equations since in conjunction with the condition for the equilibrium of the fluorescent intensity, viz.

$$\frac{dn_j}{dt}=0=\frac{dl_j}{dt},$$

they obviously lead to the relationship $X_{0j}=k_j$ where X_{0j} is the equilibrium fluorescent intensity. Thus it is justifiable to neglect the number of photons encountering activated centres.

It is shown in Appendix D that the solutions of these equations can have no periodic term when the asymptotic values of n_j and l_j are very nearly reached, but may do so during the preceding stages of the build-up. Since the solution for n_2 in terms of n_1 ((7.1) ($j=2$)) contains n_1 integrated once, the periodic parts of the expressions for n_2 and n_1 differ in phase by a right angle. Approximately the same phase difference would be expected between the curves of n_{10} and n_{20} versus exciting intensity values corrected to the same activation time, since increasing I roughly corresponds to speeding up the build-up. The number of our observations is not, however, large enough to permit of making this test.

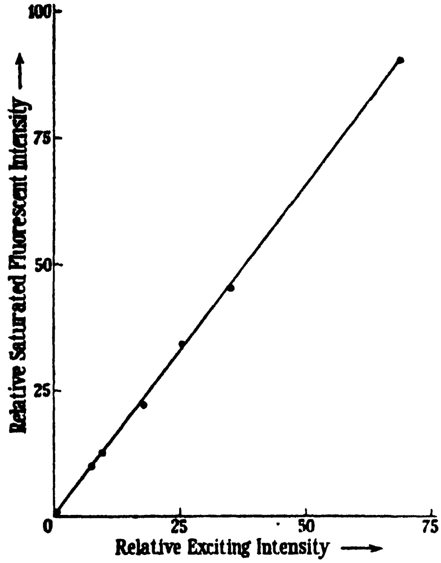


Figure 6. The variation of the saturated fluorescent intensity with intensity of exciting radiation.

If the presence of trapping centres is neglected, it is difficult to tell whether or not the solutions of the build-up equations have any periodic term. They can, however, have a maximum (Appendix E): such an effect has been observed by de Groot (1939 b). This unexpected result is given in full in Appendix E, where the equilibrium values of n_1 and n_2 (after long excitation) are also derived with and without traps.

APPENDIX A

Derivation of the series used in the two-centre decay analysis

Equation (5.8) gives

$$\beta_1 n_{10} t + C = \frac{1}{u + \zeta u^{\rho}} + \zeta(\rho - 1) S_u, \qquad \ldots\ldots(\text{A.1})$$

where $$S_u = \int_0^u \frac{u^{\rho-3}\,du}{(1+\zeta u^{\rho-1})^2} \text{ and } t = 0 \text{ when } u = 1.$$

In the integrand $(1+\zeta u^{\rho-1})^{-2}$ is expanded in powers of $\zeta u^{\rho-1}$ or $u^{1-\rho}/\zeta$ whichever <1. In the latter case, which is the more difficult, we put

$$-S_u + \int_0^\infty \frac{u^{\rho-3}\,du}{(1+\zeta u^{\rho-1})^2} = \int_u^\infty \frac{u^{\rho-3}\,du}{(1+\zeta u^{\rho-1})^2}. \qquad \ldots\ldots(\text{A.2})$$

In the integral on the left we put $\zeta u^{\rho-1} = y^2$. The resulting integral is obtained in terms of a contour integral round an infinite rectangle in the upper half-plane. Integrating the right-hand side term by term after expanding,

$$\begin{aligned} S_u &= 2\zeta^{\delta-1}\delta \int_0^\infty \frac{y^{1-2\delta}\,dy}{(1+y^2)^2} - \frac{1}{\zeta^2}\left\{\frac{u^{-\rho}}{\rho} - \frac{u^{1-2\rho}}{\zeta(\rho-\frac{1}{2})} + \frac{u^{2-3\rho}}{\zeta^2(\rho-\frac{2}{3})} - \frac{u^{3-4\rho}}{\zeta^3(\rho-\frac{3}{4})} + \ldots\right\} \\ &= \zeta^{\delta-1}\frac{\pi\delta}{\sin\pi\delta} - \frac{1}{uS(\rho-1)}\left\{\frac{u^{1-\rho}}{\zeta\left(1+\dfrac{1}{\rho-1}\right)} - \left(\frac{u^{1-\rho}}{\zeta}\right)^2 \frac{1}{1+\dfrac{2}{\rho-1}} \right. \\ &\qquad \left. + \left(\frac{u^{1-}}{\zeta}\right)^3 \frac{1}{1+\dfrac{3}{\rho-1}}\ldots\right\}. \qquad \ldots\ldots(\text{A.3}) \end{aligned}$$

Expanding the denominators, writing $\delta = 1/(\rho-1)$, collecting powers of δ and substituting in (A.1),

$$\beta_1 n_{10} t + C = \frac{\zeta^{\delta}\pi\delta}{\sin\pi\delta} + \frac{1}{u}\left\{\delta \log_e\left(1+\frac{u^{1-\rho}}{\zeta}\right) - \delta^2 z\left(2, \frac{u^{1-\rho}}{\zeta}\right) + \delta^3 z\left(3, \frac{u^{1-\rho}}{\zeta}\right)\right\}, \qquad \ldots\ldots(\text{A.4})$$

where $$z(g, x) = x - \frac{x^2}{2^g} + \frac{x^3}{3^g} - \frac{x^4}{4^g} \qquad \ldots\ldots(\text{A.5})$$

By a very similar process we obtain the expansion of S_u when $\zeta u^{\rho-1} < 1$:

$$\beta_1 n_{10} t + C = \frac{1}{u}\left\{1 + \delta\log_e(1+\zeta u^{\rho-1}) + \delta^2 z(2, \zeta u^{\rho-1}) + \sum_{g=3}^{\infty} \delta^g z(g, \zeta u^{\rho-1})\right\}. \qquad \ldots\ldots(\text{A.6})$$

z has been tabulated for all values of g when $x = 1$ (Jahnke and Emde, 1943). The values for $g = 2$, 3, 4 are respectively ·8225, ·9015, ·9470. We have calculated $z(2, x)$ for some other values of x:—

Table 3. Values of $z(2, x)$ and $\dfrac{z(2, x)}{x}$

x	0	0·2	0·4	0·6	0·8	1·0
$z(2, x)$	0	0·1908	0·3658	0·5281	0·6798	0·8225
$\dfrac{z(2, x)}{x}$	1	0·9540	0·9146	0·8802	0·8497	0·8225

With a fairly small δ graphical interpolation of $z(2, x)$ is adequate in all cases. and for the terms involving larger g's we can replace z/ζ by approximations independent of ζ, viz:

$$\sum_{g=3}^{\infty}(-1)^g\delta^g z(g,x) \doteqdot \frac{\delta^3 x}{1+\delta} \doteqdot x\left(0{\cdot}9\delta^3+\frac{\delta^4}{1+\delta}\right), \qquad \text{......(A.7)}$$

and

$$\sum_{g=3}^{\infty}\delta^g z(g,x) \doteqdot \frac{\delta^3 x}{1-\delta} \doteqdot x\left(0{\cdot}9\delta^3+\frac{\delta^4}{1-\delta}\right) \qquad \text{......(A.8)}$$

APPENDIX B

Decay laws when n is not constant throughout the layer

As mentioned in § 5, we have assumed that n does not vary from place to place in the powder at a particular instant. 60% of the crystals are less than 8 microns in diameter, which is less than 10 times de Groot's (1940) estimation of the mean path of electrons in the conduction band; since they move freely, we would expect equations (5.1) to (5.4) to be exact for single crystals in the layer, but we must allow for observing luminescence from many crystals at once, at different depths.

About 99% of ultra-violet light was absorbed and about 75% of white light. There is much scattering, so we would expect this absorption to be higher than for a single crystal. De Groot estimated that it is 10^3 cm.$^{-1}$ at 3700 A. for a single crystal. If d_0 is the thickness and λ the coefficient, we take

$$e^{-\lambda d_0} \doteqdot {\cdot}01 \quad \text{or} \quad \lambda d_0 = 4{\cdot}605. \qquad \text{......(B.1)}$$

Taking $d_0 = 0{\cdot}008$ cm. this gives $\lambda = 577$ cm.$^{-1}$ in our case. However, d_0 is probably over-estimated and the void should also reduce the absorption coefficient for a packed powder below that for one crystal.

Jesty (1946) summed intensities for a number (up to 20) of thin layers of material.

De Groot (1940) calculated the effect of summing intensity-time relationships from crystals at continuously varying depths, taking the simplest possible function of t for X, viz.,

$$X = \frac{\beta_1 n_{10}^2}{(1+\beta_1 n_{10} t)^2}. \qquad \text{......(B.2)}$$

n_{10} is assumed to vary continuously. De Groot (1940) assumed that n_{10} varied exponentially with depth. We think this is the best approximation for a packed layer of fine particles. The law is unaltered for short and for long times; at intermediate values a different law is produced. But in no case does $\frac{d}{dt}(X^{-\frac{1}{2}})$ vary with the initial value of n_1, and hence with the initial exciting intensity, so that the change of β with I cannot be explained in this way.

If, however, β, as is our case, varies with I, it must also vary with depth in the layer; if the surface is activated for a time t_0 with intensity I_0, the corresponding intensity at a depth d is $I_0 e^{-\lambda x}$, where $x = d/d_0$. We should also allow for the fact that some of the phosphorescent light is absorbed in its way through the layer. The contribution to X at a depth d should therefore be multiplied by $e^{-\lambda_b(1-x)}$, where $e^{-\lambda_b}$ is the proportion of blue light transmitted by the whole layer, and

$x = d/d_0$, d_0 being the total depth. In our case this gives $\lambda_b \doteqdot 2\log_e 2$. Using suffixes x to denote values of $\beta_1 n_{10}^2$ and $\beta_1 n_{10}$ at depth x, the complete single-centre approximation is

$$X = \int_0^1 \frac{(\beta_1 n_{10}{}^2)_x e^{-\lambda_b(1-x)}}{(C_x + (\beta_1 n_{10})_x t)^2} \left[1 + \frac{1+\rho+\dfrac{2}{\rho-2}}{(C_x + (\beta_1 n_{10})_x t)^{\rho-1}} \right] dx. \qquad \ldots\ldots(\text{B.3})$$

Since the values of $\beta_1 n_{10}$ and $\beta_1 n_{10}{}^2/A$ deduced from observations correspond to (table 1) about the same periods of activation, we can estimate from them the approximate amount that $(\beta_1 n_{10})_x$ and $(\beta_1 n_{10})_x/A$ vary for a fixed I. $\beta_1 n_{10}$ (table 3, column 9) varies by about ±20% at 20° c. over a wide range of values of I, and if $(\beta_1 n_{10})_x$ varies by this amount in (B.3) it is found that a law of the form (B.2) is a very close aproximation to (B.3) even if $\beta_1 n_{10}^2$ varies much more.

Approximate empirical relationships, obtained from table 3, are

$$\beta_1 n_{10} = 1{\cdot}75 + 0{\cdot}35 \cos 2\pi(\log I - 1{\cdot}4) = f_1(I), \qquad \ldots\ldots(\text{B.4})$$

$$\log_e \frac{\beta_1 n_1{}^2}{A} = 2\log_e I - 1{\cdot}5 + 0{\cdot}47 \cos(\log_{10} I - 1{\cdot}45) = f_2(I). \qquad \ldots\ldots(\text{B.5})$$

In estimating these quantities as a function of the depth, we assumed that we could put

$$(\beta_1 n_{10})_x = f_1(\theta I e^{-\lambda x}), \qquad \ldots\ldots(\text{B.6})$$

and

$$\log_e \left(\frac{\beta_1 n_{10}^2}{A} \right)_x = f_2(\theta I e^{-\lambda x}), \qquad \ldots\ldots(\text{B.7})$$

where θ is a constant fairly near to 1 that is the same for all curves at 20° c. It is equivalent to assuming that at a fixed depth $\bar{x}\left(= \frac{\log \theta}{\lambda} \right)$, $(\beta_1 n_{10})_{\bar{x}}$ and $\left(\frac{\beta_1 n^2_0}{A} \right)_{\bar{x}}$ have the values of (B.4) and (B.5) for a given I. This is obviously accurate if the main contribution to x comes from a thin slice of the layer near to the surface, which is so if $\beta_1 n_{10}^2$ increases rapidly with I. The effect needs to be investigated further, especially with more than one activating centre present.

APPENDIX C

Decay equations for ionized activating centres and trapping centres

1. It is first assumed that one trap is associated with each positive hole, which cannot be filled from the conduction band when the trap is occupied by an electron.

l_1 is the concentration of electrons trapped at centres of type 1, n_1 the number per unit volume of positive holes associated with unoccupied traps, n the concentration of free electrons. Then

$$X = -\frac{d}{dt}(n_1 + l_1 + n_2 + l_2), \qquad \ldots\ldots(\text{C.1})$$

$$n = n_1 + n_2, \qquad \ldots\ldots(\text{C.1}')$$

$$\frac{dn_j}{dt} = -(\alpha_j + \beta_j) n n_j + \gamma_j l_j (1 - p_j); \qquad \ldots\ldots(\text{C.2})$$

$$\frac{dl_j}{dt} = \alpha_j n n_j - \gamma_j l_j. \qquad \ldots\ldots(\text{C.3})$$

Various solutions have been obtained. In each case l_j in terms of n_j is given by

$$l_j = \left\{ l_{j0} + \alpha_j \int_0^t n n_j e^{\gamma_j t} \right\} e^{-\gamma_j t}. \qquad \ldots\ldots(\mathrm{C.4})$$

(i) Putting $n = n_j + N_j$ and substituting for l_j,

$$\frac{d^2 n_j}{dt^2} + \frac{dn_j}{dt}\{\gamma_j + (\alpha_j + \beta_j)N_j + 2(\alpha_j + \beta_j)n_j\} + n_j\left\{(\alpha_j + p_j)\frac{dN_j}{dt} + (\alpha_j p + \beta_j)\gamma_j N_j\right\}$$
$$+ n_j^2 \gamma_j (\alpha_j p_j + \beta_j) = 0. \qquad \ldots\ldots(\mathrm{C.5})$$

If N is changing so slowly that it can be regarded as constant, the solution is the sum of two exponentials when second-degree terms in N and N_j are negligible, viz.,

$$n_j = A_1 e^{-\lambda_1 t} + A_2 e^{-\lambda_2 t},$$

where λ_1 and λ_2 are given by

$$2\lambda = \gamma_j + (\alpha_j + \beta_j)N_j \pm \sqrt{\{\gamma_j - (\alpha_j + \beta_j)N_j\}^2 + 4\gamma_j\alpha_j(1 - p_j)N_j}. \qquad \ldots\ldots(\mathrm{C.6})$$

(ii) When N is negligible and n_1 is large there is an expansion valid at short times:

$$n_1^2 = y^2\left\{1 + \frac{A_1}{y} + \frac{A_1^3(A_2 + 2)}{24 A_2 y^3} + \ldots\right\}, \qquad \ldots\ldots(\mathrm{C.7})$$

where

$$A_1 = \frac{\gamma_1\{\beta_1 + \alpha_1(2 - p_1)\}}{2(\alpha_1 + \beta_1)^2}, \quad A_2 = \frac{\beta_1 + (2 - p_1)\alpha_1}{\alpha_1 p_1 + \beta_1}, \quad y = ae^{-\frac{\frac{1}{2}\gamma_1(\alpha_1 p_1 + \beta_1)}{\alpha_1 + \beta_1}}$$

and a is a constant.

With this expansion l_1 can be nearly proportional to n_1^2 and there is an approximate bimolecular law, but under limited conditions and for a short period.

(iii) At long times there is an exponential series satisfying (C.5) with $N_j = 0$ viz.:

$$n_j = x + \tfrac{1}{2}(B_1 - B_2)x^2 + \frac{x^3}{3!}(\tfrac{3}{2}B_1 - B_2)(B_1 - B_2)\ldots,$$

where $$x = a_1 e^{-\gamma_j t}, \quad B_1 = \frac{2(\alpha_j + \beta_j)}{\gamma_j}, \quad B_2 = \frac{\alpha_j p_j + \beta_j}{\gamma_j} \qquad \ldots\ldots(\mathrm{C.8})$$

so l_1 tends to vary linearly with n_1.

(iv) The equations are also satisfied when n_1 is a negative power of t at long times. If nn_j is a sum of negative powers of t, so that $\frac{d}{dt}(nn_j)$ is small compared with nn_j,

$$\frac{l_j}{nn_j} \sim \frac{\alpha_j}{\gamma_j}. \qquad \ldots\ldots(\mathrm{C.9})$$

Hence $$\frac{dn_j}{dt} \sim -(\beta_j + \alpha_j p_j) nn_j, \qquad \ldots\ldots(\mathrm{C.10})$$

$$\frac{dl_j}{dt} \sim 0. \qquad \ldots\ldots(\mathrm{C.11})$$

2. Case (iv) was next considered with λ_j traps per positive hole of the jth kind when it was assumed that free electrons fell into holes at a rate that did not depend on the number of associated trapped electrons.

If there are l_j trapped electrons and n_j holes per unit volume,

$$\frac{dn_j}{dt} = -\beta_j n_j n - \gamma_j p_j l_j, \qquad \text{......(C.21)}$$

$$\frac{dl_j}{dt} = \alpha_j n(\lambda_j n_j - l_j) - \gamma_j l_j, \qquad \text{......(C.22)}$$

The equation for n_j in terms of n and t, eliminating l_j, was much more complicated than (C.5), but it was found that if n and n_j were sums of negative powers of t, then

$$l_j \sim \frac{\lambda_j p_j \alpha_j}{\gamma_j} n n_j, \qquad \text{......(C.23)}$$

$$\frac{dn_j}{dt} \sim -(\beta_j + \lambda_j \alpha_j p_j) n_j^2. \qquad \text{......(C.24)}$$

3. We next consider the effect of variations of depth, cross-section and "p" in traps associated with positive holes.

Let there be $n_E dE$ holes with associated traps of depths between E and $E+dE$ of which there are $\lambda_1 n_E dE$, and $l_E dE$ trapped electrons.

Then the number of holes is

$$n_1 = \int_0^\infty n_E dE, \qquad \text{......(C.31)}$$

$$\frac{dn_E}{dt} = -\beta_1 n n_E - \gamma_E p_E l_E, \qquad \text{......(C.32)}$$

$$\frac{dl_E}{dt} = \alpha_E n(\lambda_1 n_E - l_E) - \gamma_E l_E, \qquad \text{......(C.33)}$$

$$\left.\begin{aligned} l_1 &= \int_0^\infty l_E dE, \\ n &= n_1 - l_1, \\ X &= -\frac{dn_1}{dt}. \end{aligned}\right\} \qquad \text{......(C.34)}$$

At times when $l^{-\gamma_E t}$ and $\dfrac{1}{nn_E\gamma_E}\dfrac{d}{dt}(nn_E)$ are negligible these equations are satisfied by

$$l_E = \frac{\lambda_1 \alpha_E}{\gamma_E} n_E n. \qquad \text{......(C.35)}$$

Hence

$$\frac{dn_E}{dt} = -(\beta_1 + \lambda_1 \alpha_E p_E) n_E n \qquad \text{......(C.36)}$$

and

$$\frac{dn_1}{dt} = -\beta_1 n_1 n - \lambda_1 n \int \alpha_E n_E p_E dE. \qquad \text{......(C.37)}$$

4. *Traps independent of centres (de Groot type) of one depth.* With two kinds of activating centres the number of free electrons is

$$n = n_1 + n_2 - l, \qquad \text{......(C.41)}$$

$$\frac{dl}{dt}=\alpha n(L-l)-\gamma l, \qquad \text{......(C.42)}$$

$$\frac{dn_1}{dt}=-\beta_1 n_1 n, \qquad \text{......(C.43)}$$

$$\frac{dn_2}{dt}=-\beta_2 n_2 n. \qquad \text{......(C.44)}$$

Substituting for l from (C.43) in (C.42) leads to

$$\frac{d^2n_1}{dt^2}-\frac{1}{n_1}\left(\frac{dn_1}{dt}\right)^2\left(1+\frac{\alpha_1}{\beta_1}\right)+\frac{1}{n_1}\frac{dn_1}{dt}\{\beta_1 n_1+\beta_2 n_2-\alpha(n_1+n_2)+\underline{\gamma+\alpha L}\}$$
$$+\underline{\beta_1\gamma(n_1+n_2)}=0, \qquad \text{......(C.45)}$$

and there is an analogous equation with suffixes 2 and 1 interchanged.

It is easily verified that the terms underlined can ultimately become the largest. (5.1) and (5.2) are then obtained with $\beta_1 R$, $\beta_2 R$, in place of β_1 and β_2. $R=\dfrac{1}{1+\alpha L/\gamma}$, and the remaining terms are of the order of $1/t^2$.

5. ***De Groot type traps of different depths.*** Let there be $L_E dE$ traps with depths between E and $E+dE$, with $l_E dE$ electrons trapped and escaping at a rate $\gamma_E l_E dE$. Then

$$\frac{dl_E}{dt}=\alpha_E n(L_E-l_E)-\gamma_E l_E. \qquad \text{......(C.51)}$$

(C.43) is unchanged: substituting from it for n,

$$\frac{dl_E}{dt}+l_E\left\{\gamma_E-\frac{\alpha_E}{\beta_1 n_1}\frac{dn_1}{dt}\right\}=-\frac{\alpha_E L_E}{\beta_1 n_1}\frac{dn_1}{dt}. \qquad \text{......(C.52)}$$

Then if $l_E=l_{E0}$ at $t=0$,

$$l_E=e^{-\gamma_E t}n_1^{\frac{-\alpha_E}{\beta}}\left\{l_{E0}n_{10}^{\frac{-\alpha_E}{\beta}}+L_E\int_0^t e^{\gamma_E t}\frac{d}{dt}\left(n_1^{\frac{-\alpha_E}{\beta}}\right)dt\right\}. \qquad \text{......(C.53)}$$

By integrations by parts,

$$\frac{\gamma_E}{L_E}\left\{l_E e^{\gamma_E t}n_1^{\frac{-\alpha_E}{\beta}}-l_{E0}n_{10}^{\frac{-\alpha_E}{\beta}}\right\}=e^{\gamma_E t}\left\{\frac{d}{dt}\left(n_1^{\frac{-\alpha_E}{\beta}}\right)-\frac{1}{\gamma_E}\frac{d^2}{dt^2}\left(n_1^{\frac{-\alpha_E}{\beta}}\right)\right\}$$
$$-\frac{d}{dt}\left(n_1^{\frac{-\alpha_E}{\beta}}\right)+\frac{1}{\gamma_E}\frac{d^2}{dt^2}\left(n_1^{\frac{-\alpha_E}{\beta}}\right)+\frac{1}{\gamma_E}\int_0^t e^{\gamma_E t}\frac{d^3}{dt^3}\left(n_1^{\frac{-\alpha_E}{\beta}}\right)dt. \qquad \text{......(C.54)}$$

So if the intensity $\dfrac{-dn_1}{dt}$ is a negative power of t, when $e^{-\gamma_E t}$ and $\dfrac{1}{\gamma_E t}$ are negligible,

$$l_E\sim\frac{\alpha_E L_E}{\gamma_E\beta_1 n_1}\frac{dn_1}{dt}, \qquad \text{......(C.55)}$$

or

$$\frac{l_E}{n}\sim\frac{\alpha_E L_E}{\gamma_E}. \qquad \text{......(C.56)}$$

Thus the ratio of concentrations of trapped electrons to free electrons tends to $\int\dfrac{\alpha_E L_E\,dE}{\gamma_E}$ integrated over the whole distribution of depths, whether the numbers of free electrons per unit volume is n_1-l, n_1+n_2-l, or $n_1+n_2+n_3-l-l_\infty$ (l_∞ being the number in deep traps).

The most general equations satisfied by our observations are, therefore,

$$\frac{dn_j}{dt} = -\frac{\beta'_j n_j n}{1+\lambda_3\int\frac{\alpha_3 n_{3E}\,dE}{\gamma_{3E}}},$$

where $$\int n_{3E}\,dE = n_3 \doteqdot n_{30} \quad\text{and}\quad j=1 \text{ or } 2. \qquad \text{......(C. 57)}$$

APPENDIX D

Conditions for a periodic term in the build-up equation with more than one type of activating centre and associated trapping centres

We shall use the notation of equations (C.5) and remove the suffixes, which will all be alike. n is now the number per unit volume of ionized activating centres of the jth kind and $n+N$ is the number of electrons in the conduction band. Then eliminating l_j between (7.1) and (7.2) with this notation gives

$$\frac{d^2n}{dt^2} + \frac{dn}{dt}\{\gamma + (\alpha+\beta)(N+2n)\} + n\left\{(\alpha+\beta)\left(\gamma N + \frac{dN}{dt}\right) - \gamma\alpha N(1-p)\right\}$$
$$+\gamma n^2(\alpha p+\beta) = k\gamma(1-p+pq).$$

Let $n = y_0 + y_1$, where y_0 satisfies (D.1) when N is constant. If we put $N = N_0 + \tau$, where τ is a small function of t, and neglect second-degree terms in y_1 and τ, y_1 satisfies

$$\frac{d^2y_1}{dt^2} + \frac{dy_1}{dt}\{\gamma + (\alpha+\beta)(N_0+\tau+2y_0)\} + y_1(\alpha p+\beta)\gamma(N_0+\tau+2y_0)$$
$$+(\alpha+\beta)\left(2\frac{dy_0}{dt} + \frac{d\tau}{dt}\right) = -y_0\left\{(\alpha+\beta)\frac{d\tau}{dt} + \tau(\alpha p+\beta)\right\}. \qquad \text{......(D.2)}$$

The same equation could hold during the decay, but with a different y_0. It may be reduced to a standard form by a process explained in many text-books (e.g. Levy, *Numerical Studies in Differential Equations*, §§ 18, 18.1). If the equation is

$$\frac{d^2y}{dt^2} + P\frac{dy}{dt} + Qy = R,$$

where P, Q and R are functions of t only, we put $y = uv$, where $u = e^{-\frac{1}{2}\int P\,dt}$, so that

$$\frac{d^2v}{dt^2} + vD(t) = R,$$

where $$D(t) = Q - \tfrac{1}{2}\frac{dP}{dt} - \tfrac{1}{4}P^2. \qquad \text{......(D.3)}$$

Solutions are generally in terms of solutions with $R=0$. We can then compare (D.3) with corresponding equations with $D(t)$ constant, and may expect a periodic type of solution when $D(t)$ is positive and not when it is negative. For equation (D.2),

$$D(t) = \tfrac{1}{2}(\alpha+\beta)\left\{2\frac{dy_0}{dt} + \frac{d\tau}{dl}\right\} - \tfrac{1}{4}\{\gamma - (\alpha+\beta)(N_0+\tau+2y_0)\}^2$$
$$-\gamma\alpha(1-p)(N_0+\tau+2y_0). \qquad \text{......(D.4)}$$

During the build-up $D(t)$ will not be positive when the asymptotic values of l and n are nearly reached, but may be so at earlier stages, when dy_0/dt and $d\tau/dt$

must be positive most of the time. During most of the decay it is almost certainly negative.

APPENDIX E

Approximations to the build-up equations after long times of activation.

The equations are

$$\frac{dn_1}{dt} = k_1 - \beta_1 n_1(n_1 + n_2), \qquad \text{......(E.1)}$$

$$\frac{dn_2}{dt} = k_2 - \beta_2 n_2(n_1 + n_2). \qquad \text{......(E.2)}$$

The equilibrium values are easily found by putting

$$\frac{dn_1}{dt} = \frac{dn_2}{dt} = 0.$$

In the most general case, from equations (7.1), (7.2),

$$\operatorname*{Lt}_{t\to\infty} n_j = M_j = \frac{K_j}{\alpha_j p_j + \beta_j}\left\{\sum_i \frac{K_i}{\alpha_i p_i + \beta_i}\right\}^{-\frac{1}{2}}, \qquad \text{......(E.3)}$$

where

$$K_j = k_j(1 - p_j + p_j q_j).$$

With no trapping centres,

$$M_1 = \frac{k_1}{\beta_1}\left(\frac{k_1}{\beta_1} + \frac{k_2}{\beta_2}\right)^{-\frac{1}{2}}, \quad M_2 = \frac{k_2}{\beta_2}\left(\frac{k_1}{\beta_1} + \frac{k_2}{\beta_2}\right)^{-\frac{1}{2}}. \qquad \text{......(E.4)}$$

We can now write

$$\left.\begin{aligned} n_1 &= M_1 - a_{11}e^{-\lambda_1 t} - a_{21}e^{-2\lambda_1 t} - a_{31}e^{-3\lambda_1 t}, \\ n_2 &= M_2 - a_{12}e^{-\lambda_1 t} - a_{22}e^{-2\lambda_1 t} - a_{32}e^{-3\lambda_1 t}. \end{aligned}\right\}. \qquad \text{......(E.5)}$$

Equating coefficients of $e^{-\lambda_1 t}$ gives the equations

$$a_{11}\{\lambda_1 - \beta_1(2M_1 + M_2)\} = a_{12}\beta_1 M_1, \qquad \text{......(E.6)}$$

$$a_{11}\beta_2 M_2 = a_{12}\{\lambda_1 - \beta_2(M_1 + 2M_2)\}. \qquad \text{......(E.7)}$$

Eliminating a_{11} and a_{12}, λ_1 must be the smallest root of

$$\lambda_1^2 - \lambda_1\{\beta_1(M_2 + 2M_1) + \beta_2(M_1 + 2M_2)\} + 2\beta_1\beta_2(M_1 + M_2)^2 = 0. \qquad \text{.....(E.8)}$$

The roots are positive and real, and we can write

$$2\lambda_1 = \beta_1(2M_1 + M_2) + \beta_2(M_1 + 2M_2) - \sqrt{M_1^2(2\beta_1 - \beta_2)^2 + M_2^2(\beta_1 - 2\beta_2)^2 + 2M_1M_2(2\beta_1^2 - 3\beta_1\beta_2 + 2\beta_2^2)}. \qquad \text{......(E.9)}$$

If λ_2 is the other root, the solutions will be sums of exponentials when indices are linear combinations of λ_1 and λ_2. For large enough values of t all but the first exponential terms will be negligible, so that

$$n_1 \doteqdot M_1 - a_{11}e^{-\lambda_1 t}, \qquad \text{......(E.10)}$$

$$n_2 \doteqdot M_2 - a_{12}e^{-\lambda_1 t}. \qquad \text{......(E.11)}$$

From (E.6) and (E.4)

$$\frac{a_{11}}{a_{12}}\beta_2 M_2 = \tfrac{1}{2}\beta_1 M_1\left(2 + \frac{k_2\beta_1}{k_1\beta_2} - \frac{\beta_2}{\beta_1} - \frac{2k_2}{k_1}\right) - \tfrac{1}{2}\sqrt{M_1^2(2\beta_1 - \beta_2)^2 + M_2^2(\beta_1 - 2\beta_2)^2 + 2M_1M_2(2\beta_1^2 - 3\beta_1\beta_2 + 2\beta_2^2)}, \qquad \text{......(E.12)}$$

which is negative if $\beta_2 > 2\beta_1$, and probably in other cases also.

Thus either n_1 or n_2 rises above its equilibrium value. If $\beta_2 \gg \beta_1$ and k_1 and k_2 are of the same order of magnitude, the solution for n_1 will not differ greatly from that with n_2 negligible, so n_2 must then have a maximum.

§ 8. CONCLUSION

The emphasis in this work has been on the detailed investigation of the time variation of the luminescence over a limited period rather than a general investigation over long periods. Even so, the analysis of the observations assuming a bimolecular law and two kinds of activating centre indicates that for quantitative work on the first few milliseconds of the decay yet more detailed observations are required during that period.

The idea that a phosphor may contain more than one kind of activating centre is not new (Martin and Headrick (1939)), but as far as is known no work has previously been done on analysing decay curves into components due to the various centres when electrons are excited into the conduction band. For our phosphor, during fluorescence and the very early stages of the decay, the contribution from the silver centres is most important; during the intermediate stage of the decay, the contribution from zinc centres predominates and at long times the emission is mainly due to copper centres. However, each centre has some effect at all times.

The variations with temperature and exciting intensity of the constants obtained by the two-centre analysis have led to a new conception of the trapping mechanism and to the conclusions that most of the electrons are in traps associated with copper centres and that during the period of our observations retrapping is important. We have also shown that these conclusions do not depend on the nature of the trap-depth distribution and that the presence of deep traps of other kinds is not excluded.

The number of our decay curves is not sufficient to justify quantitative work on the variation of the constants. It is interesting to note that variations in conditions at different depths in the phosphor are more or less compensated for in our case by variations with I of the constants. This probably does not happen in general, and so for thick layers of other phosphors the decay laws may be very complicated (e.g. Jesty (1946)).

It is thus seen that the nature of the time variation of the phosphorescence depends on many factors. We have neglected the fact that our observations correspond to intermittent excitation (ratio excitation : decay periods : : 1 : 18) and have only mentioned hole migration. In studying the time variation of the fluorescence there are even more factors to take into account, and so little has been done on this.

Finally, it can be said that although we modify the simple bimolecular theory considerably, it remains the basis of our interpretation.

ACKNOWLEDGMENTS

The authors are indebted to Mr. R. S. Longhurst for assistance with the calculations and to Mr. W. Ramsden for measuring the spectral distribution. They also thank Dr. H. A. Klasens for stimulating criticisms and suggestions, and Mr. J. A. M. van Moll and the directors of Philips' Lamps Ltd. for permission to publish this paper.

REFERENCES

Beese, 1939. *J. Opt. Soc. Amer.*, **29**, 26.
Blokhinzev, 1937. *Phys. Z. Soviet Union*, **12**, 586.
Fonda, 1945. *Trans. Electrochem. Soc.* Preprint 87–9.
Garlick and Wilkins, 1945. *Proc. Roy. Soc.*, A, **184**, 408.
de Groot, 1939 a. *Physica*, **6**, 275.
de Groot, 1939 b. *Physica*, **6**, 393.
de Groot, 1940. *Physica*, **7**, 432.
Gudden and Pohl, 1921. *Z. Phys.*, **4**, 206.
Gurney and Mott, 1937. *Proc. Phys. Soc.*, **49** (extra part), 32.
Henderson, 1939. *Proc. Roy. Soc.*, A, **173**, 323.
v. Hippel, 1936. *Z. Phys.*, **101**, 680.
Jahnke and Ende, 1943. *Tables of Higher Mathematical Functions*, p. 323.
Jesty, 1946. Lecture delivered at the Research Laboratories of the General Electrical Company, 15 June
Johnson, 1939. *J. Opt. Soc. Amer.*, **29**, 387.
Kitchener, 1939. *Trans. Faraday Soc.*, **35**, 97.
Klasens, 1946. *Nature, Lond.*, **158**, 306.
Klasens and Wise, 1946. *Nature, Lond.*, **158**, 483.
Landau, 1933. *Phys. Z. Soviet Union*, **3**, 664.
Leverenz and Seitz, 1939. *J. Appl. Phys.*, **10**, 479.
Lewschin and Romanowsky, 1934. *Phys. Z. Soviet Union*, **5**, 379.
Martin and Headrick, 1939. *J. Appl. Phys.*, **10**, 116.
Milner, 1939. *Trans. Faraday Soc.*, **35**, 101.
Mott, 1937. *Nature, Lond.*, **139**, 951.
Randall and Wilkins, 1945 a. *Proc. Roy. Soc.*, A, **184**, 365.
Randall and Wilkins, 1945 b. *Proc. Roy. Soc.*, A, **184**, 390.
Rees, 1942. *Ann. Rep. Chem. Soc.*, **39**, 76.
Riehl, 1939. *Trans. Faraday Soc.*, **35**, 135.
Seitz, 1938. *J. Chem. Phys.*, **6**, 454.
Seitz, 1939 a. *Trans. Faraday Soc.*, **35**, 83.
Sietz, 1939 b. *Trans. Faraday Soc.*, **35**, 98.

ADDENDUM TO DISCUSSION

on the paper by R. F. Schmid and L. Gerö entitled "Photochemical decomposition of CO" (*Proc. Phys. Soc.*, **58**, 701 (1946)).

Dr. J. G. Valatin*. The comments by Schmid and Gerö on the interpretation of the photochemical decomposition of CO seem to be at variance with the views of Gaydon, but they do not disagree with available direct experimental evidence on the absorption spectrum. As is emphasized by Schmid and Gerö, they conclude from the observed predissociation effect on the $A^1\Pi$ state that the effective absorption region of the continuum at 77497 cm.$^{-1}$ is very narrow, and the transition probability is small ; it can give rise to a considerable effect in the case of a suitably chosen and concentrated, monochromatic light source, but the effectiveness of the continuum cannot be judged from plates taken with low resolving power. The spectrograms of Leifson can give no evidence at all on the effect in question in the region of the xenon line.

The three lines of evidence given in the paper of Schmid and Gerö show that the absorption of the 1295 A. xenon line cannot be due to the lines of the Fourth Positive band system. Quite apart from the discussion of an eventual larger overall width of these lines, the lines of the Fourth Positive band system belong in the corresponding spectral

---

* Since both authors are now dead, Dr. Valatin has asked if he may reply to the questions raised by Dr. Gaydon in the discussion which is printed immediately after the paper.

region to initial states of the CO molecule which are not present at ordinary temperature.

As to the dissociation energy of CO, the discussion still holds. (*See also* Long and Norrish, 1946 ; Valatin, 1946 ; Edlén, 1947.) The interpretation of the photochemical decomposition of CO gives further support to the dissociation scheme of Schmid and Gerö. The value of 170 kcal./gm.-atom deduced for the heat of sublimation of carbon into ^5S atomic states agrees well with the dynamic experiments, while the lower value of the equilibrium measurements can be explained by primary formation of ^5S atoms and by secondary processes. Corresponding calculations, resulting also in vapour-pressure curves which are in good agreement with the equilibrium measurements in the carbon arc, have been given by Schmid and Gerö (1937). Papers dealing with the thermochemical side of the question are in preparation.

REFERENCES

Edlén, B., 1947. *Nature, Lond.*, **159**, 129.

Long, L. H., and Norrish, R. G. W., 1946. *Nature, Lond.*, **157**, 486 ; **158**, 237.

Schmid, R., and Gerö, L., 1937. *Mitteilungen der berg- und hüttenmännischen Abteilung der Universitat, Sopron, IX.*, 173.

Valatin, J. G., 1946. *Nature, Lond.*, **158**, 237.

OBITUARY NOTICES

Sir JAMES JEANS, O.M., F.R.S.

Sir James Jeans, famous alike as mathematician, astronomer and mathematical physicist, and world-famous as an expositor of all three sciences, died at his home, Cleveland Lodge, Dorking, on 16 September 1946 of coronary thrombosis. He had had heart attacks a year or so previously, and had had to reduce his activities, but otherwise he had been in reasonably good health ; his last few hours were passed in intense pain.

Jeans made outstanding contributions to theoretical physics on the one hand and to astronomy and cosmogony on the other hand. In each field of thought he solved some of the most difficult problems of the day. But he was not only gifted as an investigator ; he was also superbly gifted as a writer. Besides his text-books, and his two treatises, *The Dynamical Theory of Gases* and *Astronomy and Cosmogony*, he wrote two masterpieces : the one, his Physical Society Report of 1914, *Report on Radiation and the Quantum Theory*, a gem of economical exposition which ranks with Eddington's Physical Society Report of 1918, *Report on the Relativity Theory of Gravitation*, as having substantially influenced the general acceptance of a new and fundamental physical theory ; the other, his Adams Prize Essay of 1917, published in 1919 under the title *Problems of Cosmogony and Stellar Dynamics*, which unfolds in thrilling style, but with full mathematical detail, the classical researches on the stability of forms of equilibrium of rotating masses to which Jeans himself made the dominant original contributions. These works would have sufficed for any ordinary man. But in 1928, almost suddenly, feeling perhaps that his best original work was finished, Jeans turned from research to popular exposition, and at once attained a justly-deserved success. In a series of volumes, showing no signs of the speed with which they must have been composed, Jeans traversed the ground of his own and others' researches in astronomy and physics, covering nebular, stellar and planetary evolution, thermodynamics, atomic theory, relativity and quantum theory in a fresh and engaging style, illustrating the varying orders of magnitude of astronomical and atomic quantities with many a vivid simile. Whether it were technical mathematics, detailed account of an original theory, popular astronomy, popular physics or popular philosophy, Jeans hardly wrote a dull sentence. And I am tempted to put alongside his two technical masterpieces, his semi-popular volume *Science and Music* (1938), wherein he showed a side of himself which had previously been developed only in his private life. For the great mathematician, daring speculator, modern physicist and (to be truthful) interested, but only adequate, philosopher, that Jeans in turn was, was also a lover of music, a performer on the organ, a builder of two organs at his own home and a music-room designer who thought it worth while, at the age of 60, to combine his scientific and musical knowledge for the benefit of still another public. His distinction at whatever he touched equalled his versatility. And though Jeans formed no school of research in the ordinary sense, the world of science is most emphatically the poorer by his unexpected loss.

Yet Jeans had his limitations, and to get the best out of his writings it is necessary to appreciate what these were. His limitations were those of a mathematician (I speak as a mathematician) who likes the actual problems of the universe and of the atom formulated in a tidy way, with their ragged ends all tucked in, so that it is possible to make general statements about them with possibly undue confidence that they are true. It was small wonder that in his Rede lecture, *The Mysterious Universe* (1928), he considered the Great Architect of the Universe to be a Mathematician ; that anthropomorphic and mechanical models of the universe all failing, the only elements of reality Jeans could associate with the spectacle of Nature consisted of pure thought, " the thought of one whom we must consider, for want of a wider word, as a mathematical thinker ". It is part of the same characteristic that Jeans as a physicist was happiest in dealing with the *general* problems of physics, the state of molecular chaos in a gas, the nature of the second law of thermodynamics, the equipartition of energy and the distribution of energy between matter and radiation, and that he was less successful in dealing with actual stars, interiors or exteriors, as physical objects. No one was more sure than Jeans in his grasp of physical *principles* ; but there was not the same reality about the material of which his stars were made as there was about Eddington's. The time is not yet for a comparison of these two Titans. But it may be permissible to draw attention to the fact that Eddington, originally an astronomer, and only later a physicist, had a deeper physical insight than Jeans, whilst Jeans, the mathematician, had legitimate grounds for criticizing the mathematical processes by which Eddington appeared to get some of his results. Hence their occasional clashes, and hence their widely differing conclusions about stellar constitution. But it is pleasant to be able to record that when the Gold Medal of the Royal Astronomical Society was awarded to Jeans in 1922 the presentation was made by Eddington, accompanied with one of the latter's characteristically eloquent addresses.

James Hopwood Jeans was born at Ormskirk, near Southport, on 11 September 1877, the son of William Tulloch Jeans, a parliamentary journalist ; he had two younger sisters. As a boy he was much interested in numbers ; he discovered his father's book of logarithms when he was 7 and, being unable to understand what they were for, learned the first 20 of them off by heart—and retained them in his memory till near the end of his life. He was also much interested in clocks, and wrote a short booklet, " Clocks, by J. Jeans ", at the age of 9. He went to Merchant Taylors' School from 1890 to 1896, and then entered Trinity College, Cambridge, and read mathematics. He was bracketed Second Wrangler in the Mathematical Tripos of 1898 and took a First Class in Part II of the same Tripos in 1900. He spent some time in the Cavendish Laboratory whilst holding an Isaac Newton studentship. He won a Smith's Prize in 1900 and was elected a Fellow of Trinity in 1901. About this time he had a spell of ill-health, suffering from tuberculosis of the joints, and had to spend some time in sanatoria, but he made a complete recovery. He became a University Lecturer in Mathematics at Cambridge in 1904, but from 1905 to 1909 he held a Chair of Applied Mathematics at Princeton. He returned to Cambridge as Stokes Lecturer in 1910 but resigned the post in 1912, thereafter holding no regular university appointment. He was elected Professor of Astronomy in the Royal Institution in 1935, and was annually re-elected until he resigned from ill-health in 1946.

Jeans was elected a Fellow of the Royal Society in 1906, at the early age of 28. He delivered the Bakerian Lecture, on " The Configurations of Rotating Compressible Masses ", in 1917, and was awarded that Society's Royal Medal in 1919. In the latter year he became an honorary secretary of the Royal Society, holding the post for the full period of 10 years.

He was created a Knight in 1928. He became a Research Associate of Mount Wilson Observatory in 1923. He was President of the Royal Astronomical Society for 1925–27, and President of the British Association at its Aberdeen meeting in 1934. He was given honorary degrees at a number of universities, at home and abroad. He became an Honorary Fellow of the Institute of Physics in 1929. He was awarded the Franklin Medal of the Franklin Institute in 1931. The supreme distinction of the Order of Merit was bestowed on him in 1939. He became an Honorary Fellow of Trinity in 1942.

In addition to the Bakerian Lecture of 1917, Jeans delivered, amongst others, the following formal lectures :—the Halley Lecture (1922) on " The Nebular Hypothesis and Modern Cosmogony ", the Guthrie Lecture of the Physical Society in 1923 on " The Present Position of the Radiation Problem (*Proc. Phys. Soc.*, 35, pp. 222–224, 1923), the Rouse Ball Lecture (1925) on " Atomicity and Quanta ", the Rede Lecture (1930) on " The Mysterious Universe ", the Van der Waals Lecture (1923) on " The Physical Significance of Van der Waals' Equation ", the Kelvin Lecture (1925) on " Electric Forces and Quanta ", the H. H. Wills Lecture (1928) on " The Physics ot the Universe, and the Silvanus Thompson Memorial Lecture (1931) on " What is Radiation ? "

Jeans married in 1907 Charlotte Tiffany Mitchell, daughter of Alfred Mitchell, of New London, Conn. She died in 1934, leaving one daughter. He married secondly, in 1935, Suzanne Hock, daughter of Oscar Hock, of Vienna and formerly Prague. There are two sons and a daughter of the second marriage. The first Lady Jeans gained some reputation as a poetess. The second Lady Jeans is a concert organist, and has given concerts on tours at home and abroad.

I have already mentioned Jeans' musical interests. He had an organ built for himself at his home at Dorking during the lifetime of the first Lady Jeans. When he married again, he had a second organ constructed for Lady Jeans, after an antique pattern which he called a " baroque " organ, and he had a special room built for his own organ, the two rooms being acoustically insulated from each other so that he and his wife could play without either disturbing the other. Jeans had played the organ from the age of 12, but he would never perform even before close friends. He could play the whole of Bach's organ works, and his preferences were for contrapuntal compositions (Lady Jeans played Bach at his funeral). His *Science and Music* covers an enormous range of physics, and it has been highly praised. Starting with the acoustics of the human ear, it went on to explain, for the benefit of the non-specialist, the nature of pure tones, of scales and keys, of the various musical instruments, and concluded with an account of the relation of an orchestra to a concert hall, the materials to use and the optimum size of a hall for a given orchestra. The writer remembers once falling in with Jeans at an " open " day at the N.P.L., when demonstrations in the acoustics department were being given, and being amazed at the wealth of technical knowledge about sound that Jeans had at his finger ends.

Jeans' scientific work was divided between physics and astronomy. In the latter field his main contribution was to the series of forms of equilibrium of rotating, gravitating, incompressible and compressible masses, and their stability. He finally settled the difficult problem of the stability of Poincaré's " pear-shaped " figure of a rotating liquid, showing how Sir George Darwin had been misled into considering it as stable. From the now-demonstrated instability, Jeans inferred a cataclysmic origin for double stars as produced by fission of rotating masses of stellar order ; he inferred that compressible masses would in general develop a lens-shaped figure with a sharp equatorial edge, from which matter would be ejected at two antipodal points determined by the tidal action of the rest of the universe, and he saw in this a possible origin for the forms of spiral nebulae. But in none of the effects of pure rotation could he find anything resembling the solar system, or system of one large body with much smaller bodies circulating round it. To account for the occurrence of the solar system he invoked the tidal effects of a passing star, which would raise jets in the primitive sun, these jets condensing into planets and yielding massive planets at middle distances, smaller ones further out and closer in. But he reckoned that such encounters would be excessively rare events. Jeans did much more for cosmogony than merely indulging in speculations. His Adams Prize Essay of 1919 contains the backbones of many fundamental calculations in this field, which must form the starting points of future investigations which will take into account the phenomenon of the expanding universe. Many of Jeans' specific conclusions will naturally need revision as time goes on. But *Problems of Cosmogony* is a great book. *Astronomy and Cosmogony* (1928) sums up all Jeans' original researches in cosmogony, but it carries less conviction than *Problems of Cosmogony*, and the physics in it, with its hypothesis that stars consist in parts of elements of atomic number 95 or more, is somewhat strained. It should be mentioned that this decade (1918–28) of Jeans' activities saw the publication of some 35 papers in the *Monthly Notices* of the Royal Astronomical Society, amongst which must especially be recorded his papers on " Radiative Viscosity ", in which he developed the cosmical importance of the transfer, by means of radiation, of angular momentum from one layer of a rotating body to another.

Perhaps astronomy was Jeans' favourite study ; its attractions for him are seen in both his earliest and latest original works. But there was a period, say 1902–1914, when physical investigations claimed his principal attention. In paper after paper (chiefly in the *Philosophical Magazine*) he built up methods for developing the statistical mechanics of matter, and radiation in equilibrium with matter, for establishing on as rigorous a basis as possible Maxwell's distribution law for molecular velocities in a gas, and the theorem of equipartition of energy amongst the different degrees of freedom, and for determining the partition of energy between matter and radiation. These researches treated a gas of N molecules, where N is large, as a single dynamical system. He applied similar methods to the " ether " in an enclosure, resolving the fluctuating electromagnetic field therein into its harmonic constituents, calculating its number of degrees of freedom, and finally establishing what is now known as the Rayleigh-Jeans formula for the distribution of energy, in wave-length, in black-body radiation, on the classical theory, namely, $8\pi RT\lambda^{-4}d\lambda$. Rayleigh had previously published a similar formula with a different numerical factor, but at once accepted Jeans' form. This paper of Jeans of 1905 was a culminating point. He went on working at the enigma of radiation, attempting to find classical means whereby nature avoids the " ultra-violet catastrophe " predicted by the Rayleigh-Jeans formula. Eventually Jeans came to accept the law of radiation in the form given by Planck, though he suggested modifications in the mode of its derivation. He concluded that there must be something akin to a discontinuity in Nature's conduct of the process of the interchange of energy between matter and radiation. Jeans was no uncritical acceptor of the Quantum Theory : it was only after paper on paper, trying every conceivable resource to avoid the break with classical mechanics, that he was finally converted.

He then wrote his Physical Society *Report on Radiation and the Quantum Theory* (1914), in which he incorporated the then new Bohr theory of the hydrogen spectrum. This was a superb piece of exposition. He concluded it by remarking that " the keynote of the old mechanics was continuity, *natura non facit saltus*. The keynote of the new mechanics is discontinuity ; in Poincaré's words : " Un système physique n'est susceptible que d'un nombre fini d'états distincts ; il saute d'un de ces états à l'autre sans passer par une série continue d'états intermédiaires." And he ended by a free translation of a further passage from Poincaré's *Dernières Pensées*.

Though Jeans often lectured on the fundamentals of the quantum theory after his 1914 report, he made no distinctively original further contributions ; his interests settled down to astronomy. But it is singularly appropriate that Jeans should have found Poincaré's words the most apt with which to close an epoch in his own life. For there is a very close parallel between the scientific interests of Poincaré and Jeans. Both were at once mathematicians, astronomers, physicists and philosophers; both wrote fundamental memoirs on the forms of equilibrium of rotating fluids ; both devoted much thought to cosmogony ; both were attracted by the early ideas on the quantum theory ; and both wrote explicitly popular scientific books of high literary value. Both were scientific stylists. It is perhaps as an expositor that future generations will most cherish the memory of Jeans. Whether he were writing text-book, treatise, original paper or popular volume, he was always graphic, always fluent, always (or almost always) convincing, always a stimulus to the reader's curiosity. He came to stand for modern physics and astronomy to the people at large ; and he richly deserved the full stature of that position. Physics and astronomy owe him much. But he was also an acute and tireless investigator, and it is as an investigator that his friends and colleagues will best care to remember him.

E. A. MILNE.

THOMAS HOWELL LABY, M.A., Sc.D., F.R.S.

THE death of Prof. T. H. Laby, at the age of sixty-six, brings to an end the career of one devoted to the furtherance of physics and of science generally. His efforts in this direction, often in very difficult circumstances, undoubtedly undermined his health. Despite this, he was untiring in his work, the value of which to his native country, Australia, and to the world of science in general, is still inadequately recognized.

Laby was born in Victoria, Australia, and received his early academic education at the University of Sydney. After graduating, he was awarded an Exhibition of 1851 Overseas Research Studentship, and proceeded to Emmanuel College, Cambridge, and to research work at the Cavendish Laboratory under J. J. Thomson. After a successful period there, during which he held the Joule Studentship of the Royal Society, he took up an appointment as professor of physics at Wellington, New Zealand, in 1909. This he held until 1915, when he was elected to the chair of natural philosophy in the University of Melbourne, a post he retained until his resignation in 1942. His influence on Australian physics during this period was remarkable, and it is largely due to him that Australia holds a high place in the realm of physics.

Among the many reasons why Laby played such a unique part in the development of physics in Australia was his great interest in research and the wide range of his own activities in this direction. During his tenure of the chair, there existed throughout his department an air of enthusiasm and a feeling of complete confidence in the importance of the subject, which lent a distinction apparent to undergraduates as well as research students. This led to a remarkably regular production of very keen research students—so regular, in fact, that it was a matter of great surprise if, in any year, one of the Exhibition of 1851 Overseas Studentships did not fall to a member of Laby's department.

His primary interest was in precision experimental physics, but this did not prevent him from realizing the importance of other branches of the subject. Thus he was keenly aware of the importance of theoretical physics and encouraged any students with a bent in that direction. His unusual breadth of view is exhibited by his abolition of practical examinations in the subject, despite his own special interest in experiment.

It is difficult to say in which field of precise experiment Laby was most interested ; thermal conduction, mechanical equivalent of heat, x rays, geophysics, scientific radio, all occupied his attention and were a continual source of research problems for his students and assistants. The precision determination of J by Laby and Hercus is well known, as are also the series of papers by Laby and by his assistant Kannuluik on problems of thermal conduction. Laby was actively interested in the work of the geophysical prospecting party, led by Broughton Edge in Australia in 1929, and collaborated with Edge in editing the final report of the work, which is by way of becoming a standard text-book on the subject. Besides these researches, in which he, personally, took an active share, Laby encouraged work on nuclear physics, and a neutron generator was in operation just before the War.

Among his publications the most widely used is undoubtedly the ***Tables of Physical and Chemical Constants***, compiled in collaboration with Dr. G. W. C. Kaye, and now in its ninth edition.

Apart from his academic activities, Laby played a very important part in official developments in Australian science, such as the organization of the radium supply for hospitals and the formation and operation of the Radio Research Board. As a result of the latter, Australian workers have made, and are continuing to make, very important contributions to problems of radio transmission through the atmosphere. Despite all his other interests Laby maintained a detailed knowledge of developments in radio-physics. Thus, during his visit to England in 1934, he read to the Royal Society a stimulating paper by Martyn and Pulley, and was instrumental in exciting the interest of atomic physicists in ionospheric problems. He was thoroughly convinced of the importance of physics in the development of Australian industry, and devoted a great deal of time and effort towards the often thankless and wearisome task of convincing others of this now generally accepted fact.

At the outbreak of war in 1939, there existed virtually no optical industry in Australia to meet the requirements of optical munitions supply. Laby took a leading part in the organization of the Optical Munitions Panel of Australia, of which he was the first chairman. This body was vital to the establishment of a sufficiently productive industry. Laby's real value in the war crisis cannot be measured only by this. The great contribution that Australian physicists were able to make to the defence of their country and of the British Commonwealth could not have been made if in preceding years a firm tradition of high-quality physics had not been established in Australia, largely by the efforts of the Department of Natural Philosophy at Melbourne under Laby's direction. The difficulties of doing this under conditions of isolation imposed by the great distance of Australia from

Europe and America cannot easily be over-estimated, and there is no doubt that Laby sacrificed himself unsparingly in achieving this end. H. S. W. MASSEY.

[Reprinted by permission from *Nature*, **158**, 157 (1946).]

GEORGE BLACKFORD BRYAN, O.B.E., D.Sc., M.I.E.E.

DR. G. B. BRYAN, formerly Professor of Physics, Royal Naval College, Greenwich, died at Nottingham on 29 November 1946 within a few days of his 72nd birthday. Educated at Nottingham High School and Nottingham University College, he took his B.Sc. (London) degree in 1896 with first-class honours. He gained many scholarships and prizes during this period, and from 1894 to 1896 undertook a research on electric waves on long wires in conjunction with the late Professor E. H. Barton. The results of this work were published in 1897 (*Proc. Phys. Soc.* and *Phil. Mag.*, January 1897).

These successes gained for him an 1851 Exhibition enabling him to enter St. John's College, Cambridge, for three years' research at the Cavendish Laboratory, working on the conductivity of thin liquid layers and on contact potentials under the direction of Sir J. J. Thomson. This brought him the B.A. degree by research and the D.Sc. (London).

Bryan's first appointment was as Demonstrator to the late Prof. A. M. Worthington, F.R.S., at the Royal Naval Engineering College, Devonport. In 1910 he came with Worthington to Greenwich, and on the latter's retirement in the following year he became the senior member of the staff under the Head of the Physics Department. In 1922 he was appointed to the Professorship, which he held until his retirement in 1938, in which year he was made an Officer of the British Empire in recognition of his great work for the Royal Navy. Bryan joined the Physical Society in 1916 and served on the Council from 1921 to 1926.

During the 1914–1918 war, working in co-operation with H.M. Signal School, Portsmouth, he built what was probably the first continuously evacuated triode valve and carried his investigations far enough for a 100-kw. valve to be planned with every hope of success. In 1940, and at an age when most men would have felt unable to take up new work, he responded to an urgent request to join the staff of the City and Guilds College as a Special Lecturer in order to assist in the intensive radio training of undergraduates entering the College under the Hankey scheme.

With his natural ability, Bryan's training and experience rendered him clever in devising and making apparatus for experimental work and for lecture demonstrations. His interests were mainly in the direction of applied electrical science, but he was also a successful interpreter of modern physical theories to young naval officers. He wrote and published little, not from lack of either ability or energy, but—as it seemed to his friends—from a feeling of diffidence. He did not seek publicity, and seemed to prefer serving those more in the limelight to attempting to establish a place for himself. He was a great tennis player and was good in other ball games.

One of the most loyal of colleagues, he had an exceptional power of making firm friends in all surroundings. He is sadly missed by all who knew him.

Bryan married Miss Ida Rodgers of Nottingham and he leaves one daughter.

C. L. FORTESCUE.

WILLIAM BARRON COUTTS, M.A. B.Sc.

WE regret to record the death, on 16 December 1946, at the Radcliffe Infirmary, Oxford, following an operation, of Professor W. B. Coutts, who served on the Staff of the Military College of Science from 1919 to the date of his death. Born at Kinghorn, Fifeshire, in 1885, he was educated at Edinburgh University and, after a period as a schoolmaster, was commissioned in 1915 to the R.G.A. (S.R.). He served at Gibraltar, and it was his work on the Rock which created his life-long interest in the problems of fire direction and control. In 1917 he was recalled to join the 35th Advanced Class and, after completing the course and obtaining the p.a.c.—a distinction of which he was very proud—he was appointed first Instructor and then Senior Lecturer in Range-finding at the College. In 1938 he was promoted to be Assistant Professor of Fire-control Instruments, a position which he held until his death. He was also a member of the Council of the Optical Society and for a period of five years was one of its Honorary Secretaries. He was for many years a regular contributor on optical instruments to the *Journal of Scientific Instruments*.

He was beloved by his colleagues and students, to whom he was universally known as " Willie ", and his gift of dry Scotch humour enlivened many otherwise dreary conferences and lectures. His specialized knowledge of optical and fire-control instruments and his ability to teach the subject to technical officers will be a great loss.

REVIEWS OF BOOKS

Physics and Experience, by BERTRAND RUSSELL. Pp. 26. (The Henry Sidgwick Lecture, delivered at Newnham College, Cambridge, 10th November 1945. Cambridge: The University Press.) 1*s*. 6*d*. net.

In this lecture Lord Russell addresses himself, with his accustomed clarity, to the question : " Assuming physics to be broadly speaking true, can we know it to be true, and, if the answer is to be in the affirmative, does this involve knowledge of other truths besides those of physics ? " He gives no clear-cut answer to this question, but concerns himself mainly with making clear its meaning and importance ; only the first step towards an answer is taken at the end.

The portion of physics which is assumed to be true is that in which it is thought very unlikely that any new evidence will do more than somewhat modify it ; for example, the wave theory of sound. This body of physics originated in percepts, but its constituents are very unlike the percepts which gave rise to them ; our perception of sound does not at all resemble the wave. How, then, can we acquire a knowledge of the wave from the noise ? Lord Russell holds the view that perception is the last link in an artificially limited portion of a causal chain of events which starts in a physical event and proceeds through physical space to the nerves and brain of the percipient ; the percept " is what happens when, in common-sense terms, I see something or hear something or otherwise believe myself to become aware of something through my senses ". This final event is not to be regarded as divisible into " perceiving " and an " object perceived " ; it is a single unit. The percept which we call " seeing the Sun " is describable as a bright, hot, circular something existing at the moment of perception : the corresponding physical Sun is a spherical source of complex radiation existing eight minutes earlier. The former we know directly ; the latter we infer from it. How is such an inference possible ? Incidentally, Lord Russell includes both mental and physical events in the same causal chain, but distinguishes them by the fact that mental events alone can be known by someone otherwise than by inference. The physical world therefore becomes known only by inference—indeed, it almost follows, though he does not make the deduction, that it is *definable* as the world inferred from percepts—and the problem posited—how can it be so inferred ?—stands out as a major problem of epistemology.

The one condition of true inference which Lord Russell allows himself to state is that the physical world must contain more or less separable causal chains ; for if, for example, a chain started by the Sun interfered with one started by my neighbour's wireless set—or even with one started by the Moon—I should not be able, as I am, to infer the separate characteristics of those objects in the physical world. We know, however, that such independence is not complete, for when I look at sunlight reflected from a mirror I infer my face, but when I look at sunlight reflected from a tree I infer a tree. Hence the causal chains which both started in the Sun and both ended in my percepts are traced back to different arbitrary points because of the different degrees of interference of the reflecting surface in the two cases, and if I attempt to infer the character of the looking-glass from the appearance of my face, or that of the Sun from the appearance of the tree, my inference will be at least questionable. Hence " it is clear that the relation of a percept to the physical object which is supposed to be perceived is vague, approximate, and somewhat indefinite. There is no *precise* sense in which we can be said to perceive physical objects ". One is left regretting that Lord Russell had no space to pursue the matter further because, obviously, in spite of this vagueness of relation between percept and corresponding physical object, there is, in fact, no uncertainty in my deduction about what I am seeing, and I experience no inconvenience through mistaking my face for a mirror or a tree for my face. He does add that " science consists largely of devices for overcoming this initial lack of precision on the assumption that perception gives a first approximation to the truth ", and one cannot help feeling that in the absence of an equally critical examination of those devices, the legitimacy of the preceding argument must be somewhat suspect. A line of thought which leads only to doubt of the truth of its premises must either be pursued to a final conclusion or exchanged for a more profitable one. We venture to suggest that the former alternative is impossible and the latter necessary.

What, in the most indubitable terms, can we say that the physicist does ? The answer, I think, is that he makes observations—acquires percepts—and then gives us a description of an external physical world which, if it behaved in a certain specified way, would produce those percepts. (To make discussion in a reasonable space possible I shall mean by " the physicist " someone to whom all physical observations generally accepted as trustworthy are his own percepts : we are thus not concerned with the problem of " other people "). The question then arises : should we say that the world which the physicist describes is an independently existing thing whose characteristics he " infers " from the percepts, or that it is a product of his creative imagination and formed so that it both entails the occurrence of the percepts and exhibits them as a correlated system ? There was a time when we should have said, without hesitation, that Lord Russell would choose the second mode of expression, for he has said somewhere that, as a general rule, one should speak in terms of construction rather than inference wherever possible. Here, however, he adopts the language of inference without so much as an apology, and so creates the problem with which he wrestles. He may, it is true, claim that the point in question is one on which it is not possible to speak in terms of construction, but this not only seems obviously false, but also invites the retort that it is the one question on which the rule has any importance. No one has any doubt about the appropriate language in limited considerations : Shakespeare constructed, not inferred, Caliban, and Einstein inferred, not constructed, the bending of light in a gravitational field. The difference is important only when we are considering the physical world as a whole, and if construction is not the legitimate concept there, then Lord Russell's advice loses its only important realm of application.

The question does not concern what the physicist does—on that, expressed in simple behaviouristic terms, there is probably general agreement—but rather what is the appropriate word for describing what he does. It is, nevertheless, not merely a verbal question, for if the physicist infers, the question arises, is what he infers true ? whereas if he constructs, the corresponding question is : does his construction serve his purpose ? The second question admits of an answer. It serves his purpose if it correlates his observations and does not entail anything contrary to observation ; and as new observations reveal defects in the contemporary construction he modifies it or destroys it and makes a better. This seems a faithful and adequate account of physical practice, and it is difficult to see why one need complicate it by adding the arbitrary postulate that the world-picture has a quality called " truth " or " untruth " which we must labour, without any hope of getting more than a probable answer, to determine.

The strong compulsion which many (not necessarily including Lord Russell) feel to making this addition presumably arises from the fact that in problems confined within a limited field of experience this additional quality of truth or untruth almost invariably signifies something necessary and important. If (as sometimes happens) I receive a letter stating that something which I have written is utterly absurd, I infer that the writer holds views different from mine. The inference is probably true, but it may be false ; he may, for instance, be a psychologist disinterestedly interested in observing my reaction, or one who is annoyed at having been convinced against his will. There is a meaning in saying that the inference is true or false, whether or not I am in a position to determine its truth or falsity, because independent tests are conceivable which would settle the matter ; and the possibility of those independent tests arises from the fact that experience is available outside the experience from which I made the inference. But when the question concerns the whole field of sense experience there is no independent source of information about the truth of the “ inference ”. We believe Lord Russell would hold that all that we can know about the physical world is what we can deduce from percepts. If, then, percepts exhaust their possibilities in giving us an “ inference ” without the “ truth ” label, we must necessarily remain in eternal ignorance whether that label can rightly be attached or not. In that case, what is the point of Lord Russell's problem ? It seems simpler not to bother about it.

There is another aspect of the question in which the distinction between partial and complete fields of experience is important. It is undoubtedly right to say that the physical world is brought into being in order to account for the very dissimilar world of percepts, but it is much less accurate to say that a particular element of the physical world is brought into being in order to account for a particular percept. Sound waves were not postulated to account for hearing, for, in fact, they do not account for it any better than a hypothesis of sound particles would do. Sound waves are part of a much larger body of hypotheses formed to account for a much larger field of experience, including, for example, the connection between the “ velocity of sound ” and the principal specific heats of the medium. If (as is by no means inconceivable since of the links in Lord Russell's causal chain from violin to percept the final one—from nerve disturbance to percept—is the one of which we know least, namely nothing) it should be found that the percept of hearing required sound particles rather than waves, we would not give up sound waves ; they would still be needed to account for the specific heat relation, and we should probably picture something like particles carried by waves, as was proposed in the somewhat analogous dilemma concerning light. The fact is that however smoothly the post-prandial narrative of the descent from vibrating string to sweet sound may trip off the tongue, it is the toil and sweat, blood and tears of the earlier climb that the gods exact from the philosopher, and he finds no passage from sound to string ; he must cut separate tracks from thermal, mechanical, visual, as well as auditory, percepts if he is to reach his goal.

The moral of all this seems to be that the construction or inference of the physical world from the world of percepts is such a complex matter, involving such an intricate network of connections, that the picking out of any single causal chain is highly artificial. Problems arising therefrom are arbitrary rather than inevitable. For instance, Lord Russell's original question : How can we infer from the world of percepts the very dissimilar world of physical objects ? is at least a plausible problem, but the question : How can we infer from the perceptual Sun the very dissimilar physical Sun ?, in so far as the dissimilarity of premises and conclusion is conceived as involving a difficulty, is not even that. At the first step, from percept to nerve disturbance, the resemblance is completely lost, and if there is a problem it is rather why the physical Sun ultimately arrived at should have recovered rather than lost so much resemblance to the percept. But it is hard to see what can be gained in any way by analysing the network into separate threads. All problems of physics are restricted to the physical world alone, percepts acting merely as a sort of Clerk of Works to ensure that the world is built according to specification, and all relevant problems of epistemology would seem to be concerned with the principles of inferring or constructing any kind of physical world from the whole assemblage of percepts, i.e. with the whole relation of specification to building, and not with the connection between each brick and its correlative clause. Notwithstanding Lord Russell's assumption that there is a considerable body of physics which will remain almost unchanged, we cannot neglect the possibility (I would even say probability), in view of the fate of gravitational force, light waves, eternal atoms and what not, that the present physical world will be completely transformed in the not very distant future. Our present causal chains will then lose their significance, but the general physical and epistemological problems will remain.

Physicists are not as a rule interested in the problems discussed here. Therein they are unfortunate, since they remain unaware of important implications of their achievements. Philosophers usually pay too little, but sometimes too much, respect to the work of physicists. Lord Russell is almost unique in maintaining a balanced judgment based on knowledge and understanding. The physicist who neglects what he has to offer loses much.

HERBERT DINGLE.

Methods of Mathematical Physics**, by** HAROLD JEFFREYS and BERTHA SWIRLES JEFFREYS. Pp. vii + 679. (Cambridge: The University Press, 1946.) £3 3*s.* 0*d.*

Although there have been rather a large number of books in the last few years with titles similar to this, it must not be dismissed as " just another of them." It is a book of quite exceptional importance, embodying as it does the considered ideas of one who has contributed very largely to mathematical physics. It is not specially intended for beginners, not is it a set of notes for the established worker. Rather it is a complete course, which, however, would only be intelligible to one who had had some previous mathematical training.

The first chapter is entitled the *real variable*, and takes up first the fundamental laws—associative, distributive etc.—of algebra, discusses whether the symbols in a physical equation like $s = ut$ represent numbers or physical magnitudes, including their units. Real numbers are next defined by the Dedekind *method of sections* and also by the *nest of intervals*. The remainder of the chapter is devoted to the properties of sequences, the theory of convergence (for the real variable), Riemann integration and the classical development of this theory, such as tests of convergence, mean-value theorems (including Taylor's series) and the like.

Chapter 2 gives the theory of vectors, including some applications to dynamics, and chapter 3 deals with tensors, again with applications to elasticity and hydrodynamics. In chapter 4, the algebra of matrices is developed and it is in this connexion that Rayleigh's principle is introduced. Here also we find the matrices of unitary field theory, and the reciprocal lattice used in X-ray analysis, with even a few words on integral equations.

The fifth chapter deals with multiple integrals, with of course Green's and Stokes's theorems. The problem of defining the area of a surface is presented properly, without the rash statements which appear in more elementary books.

Whereas each chapter so far mentioned deals with a subject which can be specified mathematically, without reference to applications, chapter 6 is headed *Potential Theory* and deals with the matter in a manner rather reminiscent of Routh's. It is immediately followed by a valuable chapter on operational methods. This is a subject on which the male author has already written, and the treatment here follows his earlier methods fairly closely ; the fact that this method is not identical with the method using the Fourier-Mellin theorem is stressed. The next chapter, number 8, is on physical problems soluble by operational methods, including the design problems of seismographs.

One of the most instructive chapters in the book is that on *numerical methods*, where the classical interpolation formulae are not only derived, but a really careful discussion of the relative advantages of their different forms is given. Strategems for dealing with difficult regions are not overlooked, either. Here we find the Euler-Maclaurin theorem, and also methods of solving algebraic and differential equations numerically. Perhaps it would have been asking too much to beg for the inclusion of the relaxation method as well.

With chapter 10, we return to mathematical formulation, the subject being the calculus of variations. This is seen chiefly in its dynamical aspects, Hamilton's canonical equations appearing here. It might have been conducive to clear apprehension by the student if some consideration of the relation between the calculus of variations and the theory of integral equations could have been introduced into this chapter. The following chapter, one of the longer ones, is on functions of a complex variable. With the next chapter, which treats in detail of contour integration it has most of the material found in standard treatises on the subject, except the most advanced ones. This chapter on contour integration deals also with Bromwich's integral, needed in the Fourier-Mellin treatment of partial differential equations. The general subject of complex variable is continued again in the next chapter, which takes up the subject of conformal transformation, so important in physical applications, and so fundamental in Riemann's general theory. In this particular treatise, the emphasis is strongly on the physical applications, and one is glad to note that the Joukwosky aerofoil figures here. Incidentally, there is nowhere in this book any *geometrical* treatment of inversion, sources and sinks and their superposition. It is not suggested that this is an omission ; it is rather a preference, which the reviewer shares, for analytical methods in practically every case. Chapter 14, on Fourier's theorem, differs markedly from any students' treatment known to the reviewer. Whilst the derivation of Fourier's theorem is rigorous, within the limits for which it is to be proved, it is as a whole written with more attention to applications than usual. The chapter contains Weierstrass's theorem on approximations by polynomials, and a paragraph on the detection of periodicities in experimental material.

Chapters 15 (Factorial, gamma and beta functions), 21 (Bessel functions), 23 (confluent hypergeometric functions), 24 (Legendre and associated Legendre functions) and 25 (elliptic functions) each deal with special functions of use in applied mathematics. In dealing with Bessel functions, there are some innovations of notation, whose value will be best judged when they have been considered for a few years. The confluent hypergeometric functions are naturally less used than the others mentioned, largely because of the difficulty of tabulation, but the Hh_n functions have now been tabulated, and receive their due share of attention. It is satisfactory to notice that, a general theory being less important than applicability, the Weierstrass form of the elliptic functions does not appear.

Interspersed between these chapters are others on particular mathematical topics. Thus, number 16 deals with the (general theory of) linear differential equations of the second order, and their solution by numerical methods, expansion of the solution in power series or in asymptotic expansions, and substitution of definite integrals. In chapter 17, on asymptotic expansions, there is a most useful treatment in which the method of steepest descent is fitted into perspective. A short chapter is devoted to the equations of mathematical physics, and others deal individually with wave propagation, diffusion of heat, and the applications of Bessel functions. In all these chapters, the modern methods, operational and by use of Bromwich integrals, are freely used.

From this long description, it will be seen that the book covers an enormous field, and that it has throughout a modern outlook, and puts physical applications into prominence. It is true, at the same time, that rigour is always regarded ; and there are several instances where theorems (or at least sets of conditions for the truth of theorems) are given for their physical use, which would be unfamiliar to most pure mathematicians. Each chapter is headed with a quotation, from standard literary works or elsewhere, and some of these show considerable humour. Chesterton's *Flying Inn*, for example, provides the couplet about the merry road which we did tread the night we went to Birmingham by way of Beachy Head, which is attached to the chapter on differential equations. Appended to each chapter is a selection of examples for practice. These are mostly from the Mathematical Tripos, and few of them could be called easy, but they certainly provide a means of testing the reader's grasp of the matter he has read.

Finally, there is an appendix of notes, and another containing a plea for a standardized, and new, notation for potential functions. Some will no doubt hail this gladly and others will be angered, but the present reviewer must confess to a lack of interest. The notation chosen should be clear on each occasion, but the advantages of a standard one seem to be small, more especially as we shall not cease to read papers in which other, and older, notations were used.

Many an advanced student will profit by the treatise, and so will the students of all advanced teachers who find the time to ponder the book and consider what lessons they may gain from it for use in their own teaching. J. H. A.

Piezoelectricity, by W. G. CADY. Pp. 806. (New York: McGraw-Hill Book Co. Inc., 1946.) $9.00.

The Scott Laboratory, Wesleyan University, Middletown, Conn., is probably the most active centre of research on piezo-electricity in the world, and Professor W. G. Cady, who initiated work on that subject there, still leads and inspires it. That is equivalent to saying that no physicist is better equipped for producing a standard work on piezo-electricity than is Cady, whose treatise will earn him the gratitude of all interested in crystal physics.

The book starts with a relation of the discovery of pyro- and piezo-electricity, and of the parts played by the brothers Curie, by Kelvin and by Voigt in the study of phenomena and the development of theory. Short biographical notices of J. and P. Curie and W. Voigt are given. It is characteristic of Cady that this introduction, like the rest of the book, abounds in unselfish tributes to the work of others with hardly a mention of his own great contribution. A concise yet adequate introduction to crystallography is followed by chapters on crystal physics, including elastic and dielectric properties, and a treatment of thermodynamic potentials, which are used subsequently in the exposition of the theory of piezo-electricity. Valuable formulae and tables for transformation to rotated axes are given, and indeed, wealth of data and references is a special feature of the work. From a critical survey of all available experimental data, Cady has assigned " most probable values " to the elastic, dielectric, piezo-electric and other constants of quartz, tourmaline and Rochelle salt. This evaluation must indeed have been a laborious task, but it will prove of immense value to all workers in this field. Good line diagrams of idealized Rochelle salt and quartz crystals are given, but, a trifle oddly, none of tourmaline.

In the theory of piezo-electricity given by Voigt, piezo-electric strain is expressed in terms of applied electric field. Measurements show, however, that in the case of Rochelle salt the quotient of strain to field varies greatly with temperature, as also does the permittivity; whereas if the piezo-electric strain is expressed in terms of electric charge or electric displacement, the quotient is almost independent of temperature. Although Cady remarks that experimental results taken by themselves lead to the expression of piezo-electric strain in terms of displacement, " nevertheless, one should not confuse that which is most easily measurable with that which is most fundamental ; and if the proportionality of stress with field has to be abandoned, it appears fundamentally more logical to assume proportionality with polarization than with a parameter that involves both P and E". Since, however, it is always charge rather than polarization which is observed, might it not be held that expression in terms of charge is preferable ?

As was to be expected from the author, the chapter on the piezo-electric resonator bears the stamp of thoroughness. Cady has long been interested in the effect of air-gap on response frequency, to which he devotes a considerable proportion of this and subsequent chapters, but certain discrepancies between measured and predicted effects of gap still appear to defy explanation.

The reader may find that Chapter 14, which deals with circle diagram representation of the behaviour of piezo-electric resonators, requires more concentration than the rest of the work, with the exception of the treatment of the theories of Rochelle salt, but it does provide the nearest approach to visualization of the factors involved.

Methods of determining the axes of quartz crystals are described, more space being devoted to etching and optical than to X-ray methods. Chapter 17 deals briefly with various types of cut, with reduction of thermal coefficients of frequency and with vibration patterns, whilst Chapter 19 is a useful summary of knowledge on the piezo-electric valve-maintained oscillator, with clear, concise physical explanations of stabilizer and oscillator action.

Rochelle salt is a much more highly piezo-electric crystal than either quartz or tourmaline, but unlike these its dielectric and piezo-electric properties exhibit marked non-linearity and dependence on temperature. The Seignette electrics, of which Rochelle salt is the best known member, stand in relation to quartz or tourmaline pretty much as iron to paramagnetic substances, and Valasek, Kurchatov, Fowler, Mueller and Cady have worked on interaction, polarization and other theories in an endeavour to explain observed phenomena. Although about a quarter of the work under review deals with this subject, the impression left is that the explanations are still very imperfect ; but no doubt, as with ferromagnetism which is treated in an appendix, the reason lies in the complexity of the phenomena.

In this short review it is not possible to give an account of all the aspects of crystal physics or electrical engineering treated ; suffice it to mention the piezo-optic, electro-optic and other optical effects, the atomic structure of some piezo-electric crystals, diffraction of light by ultrasonic waves, all of which are briefly treated, and descriptions of methods of measurement with some hints on production of vibrators—not many, but enough to bring home the importance of the matter in a book which does not attempt an exhaustive treatment of engineering aspects.

Several chapters are followed by lists of references, and the bibliography at the end has over 650 entries. Mention is made again of the valuable data interspersed throughout the text because wealth of information is one of the principal features of the treatise ; for this reason, and because of the excellence of the general treatment of piezo-electricity, it will be in demand in every physics and engineering library. From the outset the reader is conscious of the intense enthusiasm of the author for his subject, and can hardly fail to admire the thoroughness and elegance of treatment. The printers are to be congratulated on the presentation, whilst the paucity of misprints bears witness to the meticulous care taken by all concerned. Finally, the apt quotations introducing each of the 31 chapters provide a proof —if indeed one were needed—that scientific ability and proficiency in classical literature are in no way mutually exclusive.

P. VIGOUREUX.

Applied Mathematics for Engineers and Physicists, by L. A. Pipes. Pp. xiii+618. (New York and London: McGraw Hill Book Co. Inc., 1946.) 27*s*. 6*d*.

In the days of the giants—Maxwell, Kelvin, Rayleigh, J. J. Thomson—a physicist was a man who had learnt his mathematics first, and then turned to the application of mathematics to the material world, often taking up experimental work as well. If, then, they felt a lack of mathematical power to deal with a problem, it was because mathematics as a whole lacked the particular technique, and they were able to set about filling the gap, as Kelvin devised the method of inversion, and Rayleigh that of calculating the frequencies of a system, or as Maxwell systematized spherical harmonics and Thomson applied conformal transformation to electrostatic problems. Engineers learned through the experience of their predecessors, and, with a few exceptions, found need for little mathematics.

The position is now quite different. Physicists and engineers are trained in the laboratory, and acquire some mathematics in the earlier parts of their student careers, but usually have little time for definite study later. It seems to be quite common for them to feel the lack of adequate mathematical skill after a few years, and the reviewer can testify that many have confessed this to him. There were until recently only two ways in which this situation might be handled ; accepting it, and renouncing the hope of becoming mathematically skilled, or studying mathematics as a mathematician does, after entering on one's career as physicist or engineer. Demand, however, calls forth a supply, and books began to appear which were specially designed to fill the gap felt by practical scientists. I am not sure, but I suppose the treatises by Partington and by Mellor were among the earliest, if we except the writings of John Perry. They were designed for those who had little mathematics, and had to devote space to relatively simple integration and the like. As time went on, and science students had more mathematical background, these special textbooks could take up more advanced topics, such as convergence of sequences and integrals, or elementary matrix theory, especially when the needs of quantum mechanics brought the latter subject into prominence for physicists and even for chemists.

Books of this type, then, in which matters of mathematical technique are reviewed with special reference to the needs of scientists who are presumed to have some, but not sufficient, mathematical knowledge, have tended to increase in numbers. They have appeared in Britain and America, and in Germany, and have been written sometimes by those who are themselves noted as original workers (the Jeffreys in this country, Biot and v. Karman in U.S.A., for example) and sometimes by those who have given their efforts more completely to the problems of education (the Sokolnikoffs in U.S.A. or Houstoun in this country).

Among such books, the one by Professor Pipes of Harvard must take a very high place. The didactic skill of the author shines through the whole book, and his choice of topics is valuable indeed. After a chapter on infinite series, and one on the complex variable (not the functions of such a variable), he takes up the subject of Fourier series and then that of determinants and matrices. Each chapter is followed by a number of examples, not all of them easy, and by a list of references, which, even in these four chapters, includes the standard treatises of Goursat, Bromwich, Whittaker and Watson, and Bôcher (here spelt as Bôcker). The next topic is numerical solution of algebraic equations, and then comes the theory of linear differential equations ; here the Laplace transform is introduced, and there is a table of transforms containing 84 entries. In chapters 7, 8 and 9 the methods so far described are applied to electrical and mechanical problems. This gives the occasion for introducing the notion of normal co-ordinates (matrix methods being prominent here for transformation of variables) and the Rayleigh method of calculating natural frequencies.

In the chapter on finite differences, the theory is applied to electrical filters as well as to numerical integration. Later chapters, for which we have no space to give detailed descriptions, deal with partial differentiation and the calculus of variations, individual functions including those of Bessel and of Legendre, but not elliptic functions, and vector analysis.

Chapters 16 to 21 are devoted to the solution of the wave equation and the Laplace equation, with a chapter on complex functions interpolated at an appropriate point so that the use of conjugate functions for this purpose can be used, and with the operational method (which is not distinguished from the method of the Laplace transform) occupying the last of these chapters. The book closes with a much more difficult subject, that of non-linear oscillatory systems.

The contents have been set out in considerable detail, because whether a reader of the review desires to buy the book will depend on whether these are the subjects he wants to see treated. If he does, he may be assured that they are here well presented. J. H. A.

The Kinetic Theory of Liquids, by J. FRENKEL. Pp. xi+485. (Oxford: The Clarendon Press, 1946.) 40*s*. net.

The formulation of a mathematical theory of the liquid state has attracted the attention of many workers, especially in the years just preceding the outbreak of war. It has proved to be a task of extraordinary difficulty. While the kinetic theory of gases and many aspects of the theory of the solid state have reached a highly developed mathematical form, the theory of liquids remains still, by comparison, largely empirical and descriptive. In the older theoretical considerations, emphasis was laid mainly on the analogy between the liquid state and the state of a highly compressed gas. This followed naturally from the success of theories of the kind usually associated with the name of van der Waals, which connect directly, along a single isotherm, the properties of the liquid and the gaseous states. More recently, as a result of the detailed study of the kinetics of lattice distortion in crystals, and perhaps also because of the insight gained in the study of the order-disorder transformation in alloys, it has become clear that there exists a close analogy between the liquid state and the state of an internally distorted crystalline solid. It is from this latter point of view that Professor Frenkel develops much of the theory included in the book under consideration.

To elucidate the theory the author has collected and correlated a good deal of detailed information concerning the condensed state of matter. This is woven together and unified with the theoretical work into a very readable and, indeed, fascinating book. The mathematics employed is throughout clear and simple—consisting, in fact, in many parts of the book, of simple algebraic equations relating such quantities as the activation energy, the times of relaxation, and various diffusion coefficients. It is a satisfactory feature of the book that the degree of elaboration of the mathematics is nicely related to the nature of the assumptions upon which particular calculations are based. Besides making the book easily readable, the elimination of over detailed mathematics avoids giving a false impression of the existing state of the theory.

It is possible here to mention specifically only a few of the many topics dealt with in this book. The opening two chapters are concerned with the kinetics of the crystalline state, and deal with such matters as hole formation, lattice distortion, diffusion, and the order-disorder problem. These chapters pave the way for a general discussion of the liquid state and the kinetics of fusion. Questions are raised such as: What is the immediate cause of melting? Why does melting constitute a transition of the first order, with a discontinuity in the entropy, rather than one of the second order such as occurs in the disordering of alloys? These and many other questions are critically examined in the light of the various theories put forward in recent years by the numerous workers in this field. A chapter of special interest deals, in the same critical spirit, with theoretical work on the electrical polarization of dipolar liquids, and also with the scattering of light by liquids composed of optically anisotropic molecules. There is a chapter on surface phenomena and, finally, one dealing with the properties of high polymeric substances and their solutions.

In spite of, or perhaps on account of, the clearly provisional character of the theoretical subject matter, this book is likely to prove valuable to a wide field of research workers, both experimentalists and theoreticians. Apart from the critical discussion of modern theories, the book is of great value for the many clear descriptions of relevant observational data only to be found elsewhere widely scattered throughout the literature, much of it in the Russian journals. One minor criticism should, perhaps, be made. The value of most scientific text-books, and certainly one of this kind, is greatly enhanced by a detailed index of the subject matter. It is a pity, therefore, that the present volume contains nothing more useful, in this respect, than a name index. H. JONES.

Physical Science in Art and Industry, by E. G. RICHARDSON. Second Edition. Pp. xi + 299, with 77 illustrations. (London: The English Universities Press Ltd., 1946.) 15*s.* net.

In reviewing a second edition, it is customary to refer in detail to such alterations and additions to his earlier book as the author may have made, and to congratulate him upon the evident appreciation of his efforts. These things may well be taken for granted : the writer has enriched his pages, and is entitled to a warm welcome on that account.

The previous arrangement still holds ; there are chapters dealing with locomotion, " out-of-door " physics, applications to the Fine Arts, and so forth. It is remarkable how stimulating it all becomes under Dr. Richardson's guidance. It is as if one was being taken a grand " tour du propriéteur " around the farm, along the river bank, down the mine, back via the Art Gallery—with a glimpse at contemporary architecture—finishing up in the kitchen. (This is not the exact order in the book, but all these items are there.) Now the point is this : these subjects reveal quite naturally how much applied physics have helped them both to progress and to conserve. It would be unreasonable to expect completeness in such a programme ; what is noteworthy is how little of real importance is omitted.

Readableness is not secured by superficiality. The treatment of the hysteresis loop for soils (*p*H against moisture-content), and much the same for wood, are examples of how these quite advanced conceptions can be simply explained. To what extent they are fully grasped and used in the appropriate quarters is a different matter. Another instance is a brief account of air-conditioning, a subject upon which the wildest opinions are sometimes heard—the outcome of mental confusion about what the process is intended to do. Nobody could fail to understand the principles in this case. Like all efforts of a missionary nature, there is a finite risk that the reader will go on his way rejoicing, in the belief that he has been given the " know-how " and that all the rest is easy. He may soon discover his mistake, but he must not blame Dr. Richardson, whose skill in explanation may have been a trifle intoxicating. With this caution, it is more than likely that somebody may be truly inspired by working through this book carefully, and pondering upon its implications.

At the same time, the philosophical reaction to all this activity needs consideration, namely the impact of these applications upon physics itself. Probably the ever-increasing stress of circumstances, and the urge towards increased technical efficiency (especially in " Physics down the mine ") will produce advances into the border-country between physics and engineering, and " sideways " along with physical chemistry.

If in fact there is no time like the present, then the appearance of an extended version at this moment of a first-class piece of work is an event to be welcomed, and used with a full sense of responsibility. F. IAN G. RAWLINS.

The Theory of Functions of Real Variables, by LAWRENCE M. GRAVES. Pp. x + 300. (New York and London: McGraw Hill Book Co. Inc., 1946.) 20*s.*

The subject of " real variables " can mean anything from those parts of elementary calculus which can be dealt with intuitively up to the subject dealt with by Littlewood and by Hobson in their treatises ; in general, one associates it with the more recondite theory of Fourier series, and feels that it has little of value (though maybe much of interest) for the applied mathematician.

This book, by the professor of mathematics at Chicago, sees the subject in a different light. The general view is that of the pure mathematicians named above, but it is written for those whom the author describes as " beginning graduate students ", and considerable effort has been devoted to finding and including those results of interest to the applied mathematician.

The first chapter, on formal logic, is probably the most difficult in the book. It is followed by one on the number system, after which sets of points are considered, and then begins the "ordinary" mathematics—functions, differentiation, the Riemann integral, convergence and series, all treated by the methods of real-variable theory. In chapter 9, which deals with differential equations, are several useful theorems which show beforehand whether a differential equation will have a solution differentiable with respect to a parameter, and similar theorems. The last 125 pages are given to the general theory, mainly the properties of the Lebesgue and the Stieltjes integral, and form an excellent and relatively elementary introduction to the subject.

Whilst it would not be true to say that the book is likely to be of great direct value to the mathematical physicist, it is true that it is likely to be of great interest, and that it is well written and offers him a carefully planned survey of the present extent of its territory.

J. H. A.

Retinal Structure and Colour Vision. A restatement and an hypothesis, by E. N. WILLMER with a Foreword by W. D. WRIGHT. Pp. xii+231. (Cambridge: The University Press, 1946.) 21*s*.

This book gives concisely and objectively a great deal of information on the structure and function of the retina. But its main purpose is to develop certain new ideas on the retinal end-organs and their rôle in vision. According to the main hypothesis, in its final form, stated in an Addendum, the end-organs in the human retina are of three types: cones, dark-adapting rods (absent in the central fovea) and non-adapting or so-called day-rods. As in the usual theory, scotopic vision is mediated by the dark-adapting rods. But the assumption of three kinds of cone is rejected and the three mechanisms demanded by the phenomena of colour vision are associated with end-organs in accordance with the following scheme: "red" mechanism, cones only; "green" mechanism, cones and day-rods linked to common bipolar cells and forming hybrid receptor units; "blue" mechanism, dark-adapting rods whose spectral sensitivity curve at the intensity levels required for colour discrimination has become displaced towards the blue. It may seem that the new view is in similar difficulties to the old in that it postulates structurally indistinguishable adapting and non-adapting rods outside the fovea and structurally indistinguishable cones and day-rods within the fovea. However, this is not the place to pursue criticisms of the actual hypothesis. The presentation of the hypothesis is marred by the attempt made to develop first a simpler theory involving only two colour mechanisms served respectively by rods and cones. This leads to difficult comparisons between colour diagrams based on what is essentially a theory of dichromatic vision, with standard colour diagrams referring to a trichromatic eye. Later in the book the third mechanism is introduced, but it is not clear to what extent the earlier arguments are regarded as still valid.

Despite these criticisms Dr. Willmer has many interesting things to say. In attempting a new interpretation of colour vision from a histological standpoint, he provides much food for thought for those who approach the subject with a physical background. W. S. S.

Tables of Fractional Powers, prepared by the Mathematical Tables Project. Pp. xxx+486. (New York: Columbia University Press; Agents for U.K., Scientific Computing Service, Ltd., Bedford Square, W.C. 1, 1946.) $7.50.

Tables of x^n to cover a really considerable range in both variables would of course be impracticable, and the present volume does not attempt to deal in a comprehensive way with all the possibilities. The biggest single table in the book gives x^n when $1000x$ is any prime between 100 and 1000 and n ranges from 0·001 to 1. The next largest table gives the powers from 0·001 to 0·99 of the numbers $N/100$, where N ranges from 1 to 100, and this is supplemented by tables 1 and 2, covering the cases where x is an integer less than 11 and n is once more between 0 and 1.

There is also a table of π^n, where n is between 0 and 1, and, apart from a " miscellaneous" table, the remaining ones deal with fixed values of $n(\pm\frac{1}{2}, \pm\frac{1}{3}, \pm\frac{1}{4}, \pm\frac{2}{3}, \pm\frac{3}{4})$ and give values for $x=0$ to 10. Practically every table is to 15 decimals.

With these tables, and a calculating machine, it is possible to build up most of the results likely to be wanted. The tables where x is an integer can be combined with those where it is a fraction and those for n greater than 1 with those for smaller n. Moreover, by factorizing a number between 100 and 1000, any power of it may be formed as a product from the table which takes $1000x$ as a prime.

It is pointed out in the *Introduction*, and is perhaps worth noting, that the table of 10^n is in fact a table of antilogarithms, though it would not lend itself readily to use in logarithmic computation. Attached to the *Introduction* is a bibliography, the size of which is quite astonishing ; it contains 76 entries, of which about eight are powers of e, 14 are square or cube roots, and 22 are antilogarithms. Tables of $(1-r^2)^{\frac{1}{2}}$ are given in the bibliography, but not in the tables.

The tables are reproduced from typescript and the volume is well bound, but is printed on an unattractively dark paper.

J. H. A.

Conduction of Heat in Solids, by H. S. CARSLAW and J. C. JAEGER. Pp. vi + 386. (Oxford: The University Press, 1947.) 30*s*. net.

In case others find the publications of Carslaw on classical mathematical physics confusing, it may be useful first to list them : his first publication was *Fourier's Series and Integrals and the Mathematical Theory of Heat Conduction*, published in 1906. When this had been out of print for some time, he issued a revised edition in two volumes, of which the first, issued in 1921, was called *Introduction to the Mathematical Theory of the Conduction of Heat in Solids* ; the second volume, entitled *Introduction to the Theory of Fourier's Series and Integrals*, appeared in the same year. In 1941, with Jaeger, he published *Operational Methods in Applied Mathematics*, and now we have *Conduction of Heat in Solids*, by the same two authors.

The oldest book dealt with the theory of the flow of heat governed by the equation of diffusion, and introduced the theory of Fourier's series very much as the French pioneer himself did. The methods based on conformal transformations received little attention, but the methods of images and of sources and sinks were well treated. Although Heaviside had carried out his pioneer work, ten years were to elapse before Bromwich gave it a sound mathematical basis. In the 1921 revision, the volume on heat conduction contained much the same material as the first one, but a chapter was added giving the solutions of problems which required the new methods ; the scheme preferred was that of Bromwich, expressing the solution as a contour integral and evaluating it by the method of residues. The theory of Fourier series was based on rigorous mathematical methods appropriate to the real variable—that is, by basing the work on the theory of measure and the Lebesgue integral.

The latest version has dropped all interest in the general theory of Fourier series and reverted to the outlook of the first version, where the theory of heat conduction takes priority. It differs from the forty-year-old work by including more about the methods based on complex variables, both in the guise of transformation theory and in operational methods. In this respect, it is complementary to the other book written jointly by these two authors. Here, like most recent writers, they prefer the method (not really operational) based on the Fourier-Mellin theorem, and they give a table of functions and their transforms. The book has an appreciably more practical outlook than had the older editions. There is less interest in the flow of heat in bodies of odd shapes, or heated in unlikely ways, and more in problems where the bodies and boundary conditions can be approached in practice. One boon which follows from this more practical outlook is the addition of tables of the error function and its derivatives and integrals, and of the roots of $x \tan x - C=0$, $x \cot x + C=0$, $x\, J_1(x)/J_0(x)=C$ and $J_0(x)\,Y_0(Cx)=Y_0(x)J_0(Cx)$, each for a number of values of C.

In its old forms, the book has been a standard work for as long as the reviewer can remember. The new form is likely to serve the needs of another generation at least as well.

J. H. A.

Tables of Spherical Bessel Functions, by the Mathematical Tables Project, National Bureau of Standards, Volume 1. Pp. xxviii + 375. (New York: Columbia University Press.) $7.50.

The functions tabulated in this volume are not, strictly, Bessel functions, but the quantities $J_n(x)/x^{\frac{1}{2}}$ with a normalizing factor. They are thus the functions which are directly required in solving the wave equation in spherical, conical or spheroidal co-ordinates, for they are given for values of n equal to $\pm(N+\frac{1}{2})$ where N is an integer from 0 to 13 inclusive. As Dr. Morse says in the Foreword: "It seems a far cry from the diffusion of a search-light beam by fog to the triggering of nuclear disintegrations by neutron collisions ; from the production of "knock" in gasoline engine cylinders to the electrical oscillations of an ultra-high-frequency radio tube"; but all these phenomena depend on the wave equation, and consequently may call for these functions.

The tables are given to at least eight significant figures for x up to 10, where the interval is 0·01, and to seven figures from $x=10$ to $x=25$ (interval 0·1). In general, the functions with their (modified) differences are given, but near the zeros of the functions these would not suffice for interpolation to full accuracy, and then the product of the function and a suitable power of x is tabulated in a subsidiary table so as to remove the irregularity.

The Introduction, as usual, contains much interesting matter, including series which relate these functions to the sine and cosine integrals. J. H. A.

Introduction to Atomic Physics, by S. TOLANSKY. Second Edition, with Appendix. Pp. 351. (London: Longmans, Green and Co.) 15*s*.

In Dr. Tolansky's book the student is presented with a survey of the whole of modern physics in 350 pages. The treatment uses only the simplest mathematics, but nearly all the traditional derivations of fundamental formulae and a few new ones have been worked into the text. On the whole the treatment is too concentrated. A student who is to understand the subject matter of the book will have to bring a very good knowledge of classical physics to bear before the full implications of many of Dr. Tolansky's concise phrases are apparent to him. However, as a guide to a student who wishes to select subjects for extensive study the text is excellent. Some references to standard texts are made at the end of each chapter, and these could with advantage be increased in a further edition. On the whole the book is free from misstatements of fact, and such as there are can fairly be attributed to an attempt to gloss over difficulties. Few would agree that the electrolytic concentration of deuterium can be wholly accounted for by differences in mobility, and the historical account of the hydrogen spectrum is less than just to Rydberg's contribution to spectroscopy: the account of the determination of the range of α particles is unnecessarily old-fashioned. In speaking of nuclear isomerism, the statement that the iosmers can sometimes be separated chemically, although true, is misleading, unless more fully explained. The book concludes with a good elementary chapter on relativity, and a useful appendix on recent determinations of the atomic constants ; a second appendix summarizes the main part of the new scientific information contained in the Smyth Report. The binding and paper are reasonable and the price moderate. C. H. C.

The Vector Operator j, by F. C. GILL, A.M.I.E.E., A.I.Mech.E. Pp. 61, with 32 diagrams. (Pitman, 1946.) 7*s*. 6*d*. net.

This little book is written to explain to the mathematically uninitiated how the complex operator is used in the calculation of quantities in problems of alternating current engineering relating to single-phase, two-phase and three-phase circuits. Naturally no fundamental treatment is possible, nor is it attempted, within the limits of size adopted. The method demands care in choice of symbols and in their correct use. This correctness is usually attained in the examples selected, but a few slips occur which may cause confusion, and should be eliminated in any further edition. The book should be found helpful to engineers in the field indicated. D. O.

THE PROCEEDINGS OF
THE PHYSICAL SOCIETY

VOL. 59, PART 4 1 July 1947 No 334

SOME OBSERVATIONS OF THE MAXIMUM FREQUENCY OF RADIO COMMUNICATION OVER DISTANCES OF 1000 KM. AND 2500 KM.

BY W. J. G. BEYNON,

The National Physical Laboratory; now at University College, Swansea

MS. received 27 *September* 1946; *read* 15 *November* 1946

ABSTRACT. Measurements have been made on the magnitude of the discrepancies to be expected in the practical application of maximum usable frequency (M.U.F.) calculations. The maximum usable frequencies for radio transmission over distances of 1000 km. and 2500 km. have been deduced from continuous observations at sunrise and sunset of the relative field strength of signals received at Slough from short-wave broadcasting stations located near Berlin and Moscow respectively. It was found that the discrepancy between mean calculated and observed values of M.U.F. amounted to -3% for 1000 km. and to -11% for 2500 km. From a critical examination of the results it is concluded that a large proportion of these errors, particularly for the distance 2500 km., was due to inadequate knowledge of the ionosphere characteristics near the mid-point of the trajectory, and that the real errors of the mean calculated values probably do not exceed $\pm 2\%$ at 1000 km. and $\pm 3\%$ at 2500 km.

§ 1. INTRODUCTION

In practical radio-communication problems, an accurate knowledge of the maximum usable frequency which can be transmitted from one point to another is of major importance. Several methods of calculating this quantity from vertical-incidence ionospheric data are available (Smith, 1938; Millington, 1938; Appleton and Beynon, 1940), and curves of the maximum usable frequency (M.U.F. curves) now form part of the regular ionospheric data issued from observatories. Qualitative information obtained from the practical application of these curves indicates that there is often considerable discrepancy between these calculated values of the M.U.F. and the experimental values; in particular, it appears that frequencies considerably in excess of the values calculated from vertical-incidence data can often be satisfactorily transmitted from one point to another. In seeking for the explanation of this apparent divergence between theory and experiment it is instructive to note the simplifications and assumptions usually involved in these calculations. The following briefly summarizes the major sources of error involved in calculating and applying M.U.F. curves.

(*a*) The application of vertical-incidence ionospheric data to oblique-transmission problems usually depends directly or indirectly on two fundamental theorems due respectively to Breit and Tuve (1926) and to Martyn (1935). In the original form these theorems were stated only for the case of a plane earth and ionosphere, and for distances of transmission greater than about 600 km. Some modifications must be introduced to compensate for the effect of the curvature of the earth and ionosphere. An exact compensation would complicate the mathematical analysis too considerably, so that practical solutions are always subject to some degree of approximation.

(*b*) The calculations are normally concerned with ordinary-ray transmission, the effect of the earth's magnetic field being subsequently added as a comparatively small correction term. The magnitude of this correction will depend on many factors, such as the ratio of the frequency to the gyro-frequency, and the distance and direction of transmission.

(*c*) It is usually assumed that for that part of the trajectory which lies within the ionosphere there is no horizontal gradient of ionization.

(*d*) In applying M.U.F. curves to practical cases, the ionospheric characteristics at the mid-point of the trajectory have usually to be interpolated from the characteristics measured at a limited number of observing stations. Some part of any discrepancy between calculated and observed values of the M.U.F. may thus be the result of inadequate knowledge of the controlling ionospheric conditions.

A really accurate test of the theoretical analysis underlying the calculation of the M.U.F. will require oblique-incidence pulse experiments over long distances of transmission, together with simultaneous normal-incidence observations at one or more intermediate stations along the path. Unfortunately, under present circumstances, this is not possible, but some interesting practical information can be obtained from observations on signals received from distant short-wave broadcasting stations. The experiments described in this paper compare observed and calculated maximum usable frequencies over sender-receiver distances of 1000 km. and 2500 km. for a practical case in which normal-incidence ionospheric data are available only at one end of the transmission path. An estimate of the accuracy which is obtained over these distances should prove useful in predicting the accuracy which is normally to be expected in oblique transmission over greater distances.

§ 2. THEORETICAL CONSIDERATIONS

Further reference will now be made to three of the points noted above concerning the calculated values of the M.U.F.

(a) *Effect of the earth's magnetic field*

In all the calculations involved in the present experiments a simple approximation has been adopted. It is assumed that in east-west transmission of frequencies near 10 Mc./s. over a distance of 1000 km., the M.U.F. for the extraordinary component exceeds that for the ordinary component by 0·2 Mc./s. For the Moscow–Slough transmissions the corresponding figure is taken to be 0·3 Mc./s. These values represent the order of the separation between the maximum usable frequencies of the two magneto-ionic components deduced from a simplified theoretical consideration of the factors involved.

(b) *Controlling ionospheric conditions*

In the experiments described here, the senders are located to the east of the receiver, and vertical-incidence ionospheric data are available at the receiving site only. Now for transmission paths along a parallel of latitude, when no direct ionospheric data are available, it is customary to assume that the variation in ionospheric characteristics with longitude is the same as, or at least similar to, the diurnal variation which is observed at any fixed station at that latitude. This is probably a reasonably valid assumption when dealing with sender-receiver distances up to about 1000 km., but for longer distances it may be expected to become progressively less and less accurate. In the present case this assumption is initially made, and subsequently it is reconsidered in the light of the actual results.

(c) *Calculated values of the M.U.F.*

The experimental results have been considered in relation to the analysis for a parabolic type of layer described in papers by Appleton and Beynon (1940, 1942). The important points involved in calculating the M.U.F. by this method are briefly given below.

For a parabolic type of reflecting layer the vertical-incidence relation between equivalent height of reflection h' and frequency f can be represented by the equation

$$h' = h_0 + \frac{x \cdot y_m}{2} \log_e \frac{1+x}{1-x},$$

where

$$x = f/f^0.$$

f^0 is the ordinary-ray critical frequency, and y_m and h_0 are constants of the layer which can readily be determined from the experimental vertical-incidence (h', f) curve. When f^0, y_m and h_0 are known, the maximum usable frequency for any distance of transmission can be read from a graph. It can readily be shown that the value of the calculated M.U.F. depends principally upon the magnitude of $(y_m + h_0)$, and only to a very much lesser extent on the individual values of y_m and h_0. The vertical-incidence critical frequency f^0 can usually be measured to within ± 0·1 Mc./s., and $(y_m + h_0)$ can be determined to within ± 10 km. If f^0, y_m and h_0 are parameters of the controlling part of the ionosphere, a single calculated value of the maximum usable frequency for distances of about 1000 and 2500 km. should then be accurate to within about ± 3% and ± 4% respectively.

§3. EXPERIMENTAL PROCEDURE

The first observations were made in November 1942 on the signal received at Slough (lat. 51° 30′ N., long. 0° 33′ W.) from the Zeesen high-frequency broadcasting stations presumed to be situated at Königswusterhausen (lat. 52° 18′ N., long. 13° 37′ E.). Some of the reasons which prompted a study of transmissions from these senders are given below.

(*a*) The transmission path lies approximately in an east-west direction.

(*b*) The distance of Zeesen from Slough (990 km.) is not too large, and this made it probable that the controlling ionospheric conditions at the mid-point of the trajectory could be related to ionospheric conditions at Slough.

(*c*) The frequencies of the Zeesen senders were such as to ensure that one or more observations of the M.U.F. could be made during each sunrise or sunset period.

(*d*) At this latitude, ionospheric conditions are particularly suitable at the mid-winter period for experiments of the kind to be described below. There is a maximum diurnal change in region-F_2 ionization and a minimum occurrence of abnormal or intense region-E ionization.

The success of these initial experiments over 1000 km. prompted the extension of the observations to signals received from the short-wave broadcasting station situated near Moscow (lat. 55° 45′ N., long. 37° 37′ E.). The distance Moscow–Slough (2500 km.) is still well within the limit for single-hop transmission by region F_2, and the frequency of transmission was such as to ensure that signals from this station regularly became critical during the winter sunrise period. Observations necessarily had to be confined to the sunrise period, since it was found that in the afternoon transmissions on this frequency ceased before the frequency became critical. Observations on the Moscow signals were commenced in January 1943 and continued during the two succeeding winter periods.

For the Moscow–Slough path, the controlling point in the ionosphere is located at lat. 54° 30′ N., long. 17° 30′ E. In the absence of direct normal-incidence ionospheric data it would thus seem appropriate to use mean data deduced from normal-incidence observations at Slough (lat. 51° 30′ N., long. 0° 33′ W.) and from Burghead (lat. 57° 42′ N., long. 3° 30′ W.). We shall therefore assume initially that the longitude variation in ionospheric characteristics near the parallel of lat. 54° 30′ N. follows the mean of the diurnal variations observed at Slough and Burghead.

Measurements of the maximum usable frequency were obtained from a continuous record of the field strength of the received signal during the sunrise and sunset periods. Accurate observations were made of the time at which the strong ray transmission gave place to weak scatter signals at sunset or *vice versa* at sunrise. Normally it was quite easy to ascertain this critical time to within a minute or so, and at this instant the frequency of the transmission under observation was also the M.U.F. for that particular distance and direction of transmission. These field-strength observations were obtained automatically in the form of ink records of the voltage variations in the diode detector circuit of a commercial communication receiver. Preliminary tests indicated that the receiver was particularly free from frequency drift and that the tuned frequency remained accurately constant for many hours at a time. Even so, a careful check was maintained on the tuning of the receiver during the critical periods of the measurements. The aerial system was not calibrated, so that absolute field-strength values were not known, but the records gave relative values of the signal strength, this being all that was required. From measurements with a standard-signal generator it is estimated that the field strength of the received signal generally changed by a factor of at least 20 to 1 during the complete change from ray signal to "scattered" signal or *vice versa*. Tracings of actual records showing the transition from one type of transmission to the other are shown in figures 1 (*a*) and 2 (*a*). During the critical periods the automatic record was supplemented by aural observations of the received signal. Throughout the experiments, a careful watch was maintained on ionospheric conditions at Slough. These vertical-incidence observations consisted in visual measurements of the critical frequency of region F_2 at 15-minute intervals with a photographic (h', f) record once every 30 minutes.

§4. EXPERIMENTAL RESULTS

(a) *Zeesen–Slough results*

These observations were made during the two periods 30 November 1942 to 12 December 1942 and 26 January 1943 to 12 February 1943 on the 9·65 Mc./s. and 11·86 Mc./s. Zeesen senders. During this period a very large number of transitions from the one type of transmission to the other were observed, but for reasons which will become clear from the discussion given below, only those observations which should be associated with the normal sunrise and sunset changes in region F_2 are considered. The results are summarized in table 1, together with the calculated values of the maximum frequency. The longitude difference between the mid-point of the trajectory and that of Slough corresponds to a time delay of 28 minutes, so that the calculated values given in the table are those deduced from normal-incidence observations at Slough 28 minutes after the entry or exit of the ray signal, as the case might be. During both periods of observation, forty reliable measurements were obtained of the transition, and of this number, eight which could not be related to the vertical-incidence observations made at Slough were rejected. The term "no correlation" indicates that the M.U.F. calculated from Slough data was either stationary or varying in the wrong sense at the appropriate time with respect to the observed transition.

(b) *Moscow–Slough results*

These observations were made during three winter periods: January–February 1943, December 1943 to February 1944, and December 1944 to February 1945. During this period a total of 80 reliable observations of the critical times were obtained. Typical samples of 10 results from each period are given in table 2, together with the calculated values of the M.U.F. It will be noted that the average results in all three groups of observations show a divergence of 9 to 11 % between calculated and observed values.

§5. DISCUSSION OF RESULTS

(a) *Zeesen–Slough observations*

It has already been noted that in radio-communication problems it is only in very isolated cases that direct ionospheric measurements are available for the mid-point of a trajectory, and in the practical application of normal-incidence data to communication problems it is usually necessary to make some assumptions about the variation in ionospheric characteristics between the actual ionospheric stations. In east-west transmissions, such as those considered in this paper, we are particularly interested in the accuracy with which the variation with longitude can be deduced from the diurnal variation observed at a single fixed station, but few direct measurements have yet been made to examine this point. Schafer and Goodall (1939), however, have made some simultaneous critical-frequency measurements at Washington and Deal, U.S.A. These two stations are 300 km. apart, Deal being N.E. of Washington, and the published curves of these authors show that detailed fluctuations in the critical-frequency values are often repeated at Washington some time after occurring at Deal, but these results do not indicate accurately the time which elapses between the repetition of such fluctuations at the two stations. In the present experiments, a comparison between the field-strength records and the vertical-incidence ionosphere measurements at Slough yields some interesting information on this point.

Table 1. Observations at Slough on signals from Zeesen

Date	Freq. (Mc./s.)	Entry of ray signal	Exit of ray signal	Calculated M.U.F. (Mc./s.)	Difference calc.–obs. (Mc./s.)	Percentage difference	Estimated sender-receiver distance (km.), *vide* § 6
		(G.M.T.)					
Results for period 30 *November* 1942 *to* 12 *December* 1942							
30.11.42	11·86	0824		12·0	+0·1	+0·8	980
	11·86		1426	12·2	+0·3	+2·5	950
	9·65		1554	9·8	+0·15	+1·5	965
1.12.42	11·86	0921		11·8	−0·1	−0·8	1000
	11·86		1408	11·2	−0·7	−5·9	1080
	11·86		1422	11·0	−0·9	−7·5	1100
	9·65		1544	9·8	+0·15	+1·5	970
2.12.42	9·65	0750		9·05	−0·6	−6·2	1085
	11·86		1427	11·3	−0·6	−5·0	1065
3.12.42	11·86	0909		12·1	+0·2	+1·7	965
	11·86		1304	11·6	−0·3	−2·5	1030
	11·86		1341	No correlation			
	9·65		1505	10·05	+0·4	+4·1	930
4.12.42	11·86		1440	11·1	−0·8	−6·7	1090
5.12.42	9·65	0806		10·2	+0·55	+5·6	905
	11·86	0845		11·8	−0·1	−0·8	1000
7.12.42	9·65	0814		9·5	−0·15	−1·5	1010
	11·86	0840		11·6	−0·3	−2·5	1030
	11·86		1454	11·7	−0·2	−1·7	1015
9.12.42	9·65		1513	No correlation			
	11·86		1427	No correlation			
10.12.42	11·86	0919		11·7	−0·2	−1·7	1015
12.12.42	9·65	0852		10·1	+0·45	+4·6	920
	11·86	0946		11·4	−0·5	−4·2	1055
Results for period 26 *January* 1943 *to* 12 *February* 1943							
26.1.43	11·86		1328	10·1	−1·8	−15·1	1215
	9·65		1548	9·0	−0·65	− 6·7	1090
28.1.43	11·86	0846		10·3	−1·6	−13·4	1190
1.2.43	9·65	0820		No correlation			
	11·86	0931		11·0	−0·9	−7·5	1100
2.2.43	9·65	0801		No correlation			
	9·65	0819		8·4	−1·25	−13·0	1185
	9·65		1551	10·1	+0·45	+4·6	920
3.2.43	9·65	0840		No correlation			
5.2.43	9·65	0938		9·75	+0·1	+1·0	975
8.2.43	9·65	0759		9·4	−0·25	−2·6	1030
9.2.43	9·65	0752		9·6	−0·05	−0·5	1000
	11·86	0936		No correlation			
10.2.43	9·65	0754		8·65	−1·0	−10·3	1145
11.2.43	9·65	0733		8·2	−1·45	−15·0	1215
12.2.43	9·65	0812		No correlation			

Table 2

Samples of M.U.F. observations at Slough on signals from Moscow. Frequency of transmission 10·44 Mc./s.

Date	Time of entry of ray signal (G.M.T.)	Calculated M.U.F. (mean of Slough and Burghead ionospheric data) (Mc./s.)	Percentage error (%)
Results for period January–February 1943			
18.1.43	0700	10·5	0
20.1.43	0643	9·3	−11
21.1.43	0718	8·8	−16
9.2.43	0550	8·3	−20
11.2.43	0549	8·6	−18
16.2.43	0521	8·5	−19
19.2.43	0544	9·3	−11
23.2.43	0518	8·1	−22
25.2.43	0546	10·5	0
27.2.43	0548	10·5	0
		Average error	−11·7
		Average error of 17 calculated values in this period	−11%
Results for period December 1943–*February* 1944			
20.12.43	0628	8·7	−17
22.12.43	0655	10·1	− 5
4. 1.44	0637	9·3	−11
13. 1.44	0641	9·3	−11
19. 1.44	0655	10·6	+ 1
21. 1.44	0633	8·8	−16
25. 1.44	0626	10·3	− 1
27. 1.44	0631	9·0	−14
1. 2.44	0623	10·3	− 1
10. 2.44	0630	8·4	−20
		Average error	− 9·3
		Average error of 20 calculated values in this period	− 9%
Results for period December 1944–*February* 1945			
11.12.44	0636	9·7	− 7
19.12.44	0651	9·6	− 8
6. 1.45	0630	9·5	− 9
16. 1.45	0631	8·9	−15
20. 1.45	0616	10·2	− 2
25. 1.45	0612	9·3	−11
30. 1.45	0645	9·2	−12
3. 2.45	0615	9·3	−11
10. 2.45	0607	8·4	−20
16. 2.45	0558	9·9	− 5
		Average error	−10
		Average error for 43 calculated values in this period	−11%
		Average error for 80 calculated values in all three periods	−11%

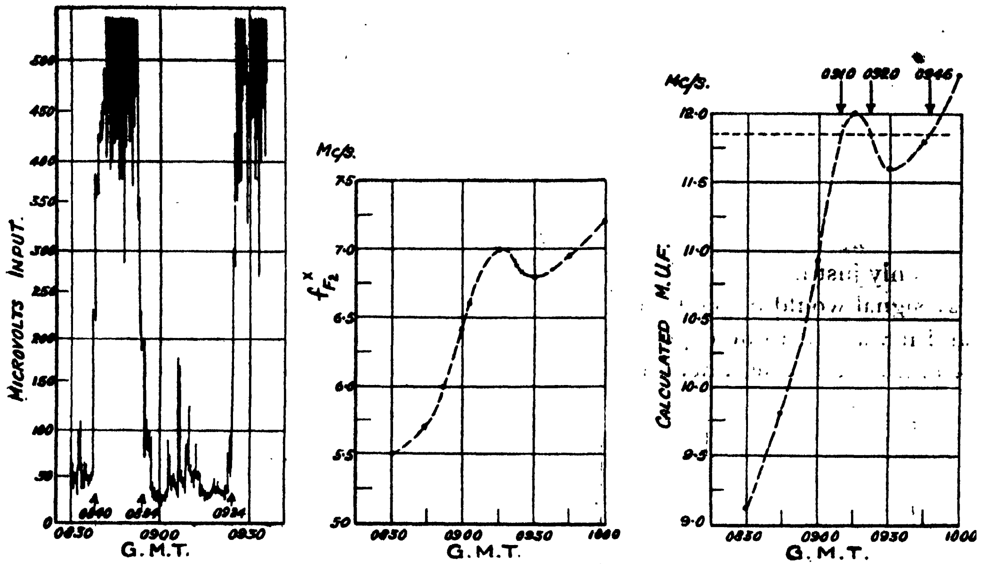
In some cases it is clear that the longitude variation in the ionosphere characteristics corresponds very closely with the local time variation observed at a fixed station, and in many other cases it is equally clear that there may be differences which are significant in relation to experiments of this kind. Figures 1 (*a*) and 1 (*b*) show an example of the detailed correspondence between the field-strength measurements on the Zeesen transmissions and the ionosphere characteristics observed at Slough. During the morning of 7 December 1942 the ray signal from Zeesen on a frequency of 11·86 Mc./s. was observed to come up sharply at 0840 G.M.T. The ray signal disappeared with almost equal abruptness at 0854 G.M.T. and reappeared again at 0924 G.M.T. The vertical-incidence critical frequency observed at Slough over this period is shown in figure 1 (*b*) and the calculated maximum usable frequency for a distance of 990 km. is shown in figure 1 (*c*). Comparing figures 1 (*a*) and 1 (*c*), it is clear that the oblique phenomena can be closely correlated with the ionosphere measurements made at the Slough end of the transmission path. The time delays between the actual entry, exit and re-entry of the ray signal and the times calculated from Slough data are 30, 26 and 22 minutes respectively. Assuming that local time and longitude variations in ionosphere characteristics are equivalent, the time delay corresponding to one-half the total transmission distance should be 28 minutes, so that in this particular example the observed time differences are quite near the expected value, but we shall see that this is not always the case. In the course of these experiments, comparatively small local irregularities in the ionosphere often caused the oblique signal to appear and disappear for short periods. On some days, particularly during the second group of observations, the signal often came up and disappeared again six or seven times during a period of a few hours. This is well illustrated in figure 2 (*a*), which shows the field-strength record of the 11·86 Mc./s. signal on 28 January 1943. The M.U.F. curve calculated from Slough data is shown in figure 2 (*b*), and, for convenience, the times at which the ray signal was present are also indicated in this figure. It will be noted that during the period of this record the ray signal appeared and disappeared no fewer than eight times. Most of these transitions can be correlated with maxima in the calculated M.U.F. values, but the time delays are clearly much larger than the 28 minutes which are normally assumed for this path length. A large proportion of the discrepancy between the calculated and observed M.U.F. for the sunrise period of this particular day (see table 1) can be ascribed to the fact that the time delay was of the order of 60 minutes rather than the assumed value of 28 minutes.

In the light of this discussion we now reconsider the observations given in table 1. The results for the period 26 January 1943 to 12 February 1943 differ from those for the earlier period in three main points:—

(*a*) The upper frequency (11·86 Mc./s.) was only received on very few occasions, and most of the observations were thus of necessity confined to the frequency 9·65 Mc./s.

(*b*) The discrepancy between calculated and observed values in this group of observations is considerably greater than that observed in the earlier period.

(*c*) There is an increased number of cases of "no correlation" with vertical-incidence Slough data.



(*a*) Recorded signal from Zeesen (11·86 Mc./s.).

(*b*) Vertical incidence measurements at Slough.

(*c*) Calculated oblique-incidence M.U.F.

Figure 1. Observations made at Slough, 7 December 1942.

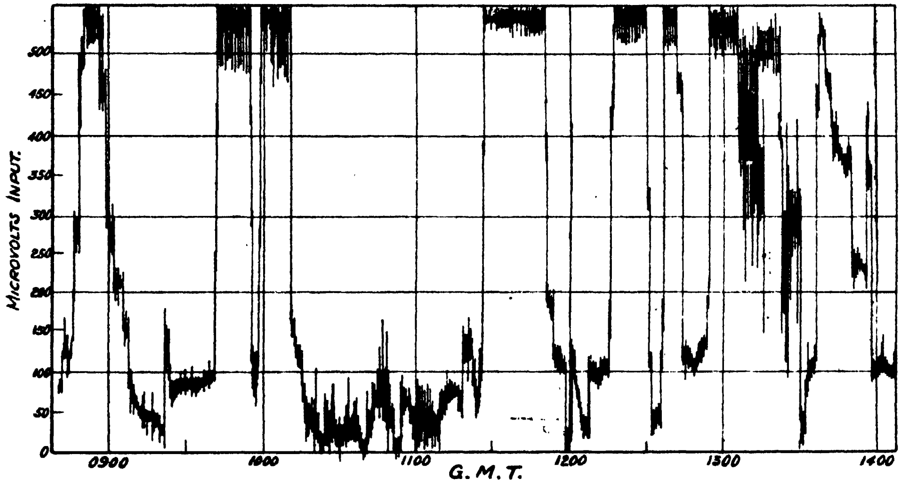


Figure 2 (*a*). Signal from Zeesen (11·86 Mc./s.) recorded at Slough, 28 January 1943.

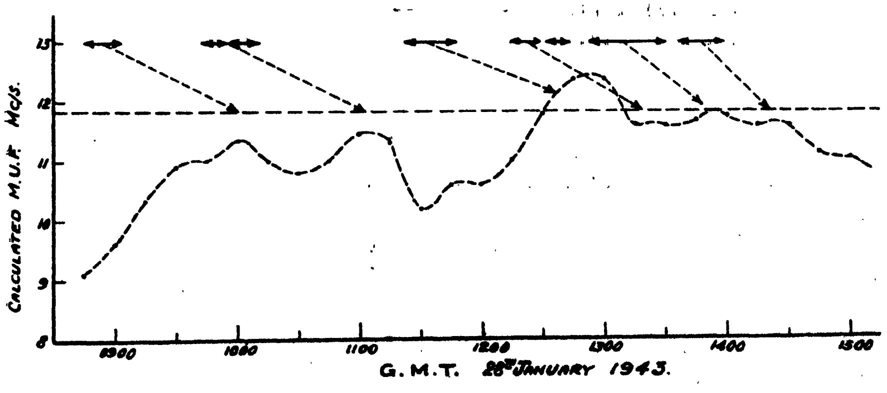


Figure 2 (*b*). Correlation of maxima in M.U.F. curve with observed 11·86 Mc./s. ray signal. Periods at which ray signal was observed are indicated thus ←——→

Figure 3 shows the mean diurnal variation of critical frequency measured at Slough for the days in each of two periods of observation. It will be noted that smaller values of critical frequency were observed during the second period. At this time, the critical frequency seldom increased sufficiently to permit observation of a ray signal from the 11·86 Mc./s. sender, and often the 9·65 Mc./s. ray signal was only just in. During such critical conditions, the reception or otherwise of a ray signal would depend on comparatively small irregularities in the ionosphere, and it was not to be expected that such small variations in ionosphere structure would be a simple function of solar angle. Now the M.U.F. calculations given here really assume that ionospheric conditions are a simple function of solar angle, and it is thus not surprising that these calculations should show the greatest error on days when the M.U.F. was only slightly larger than the frequency of communication.

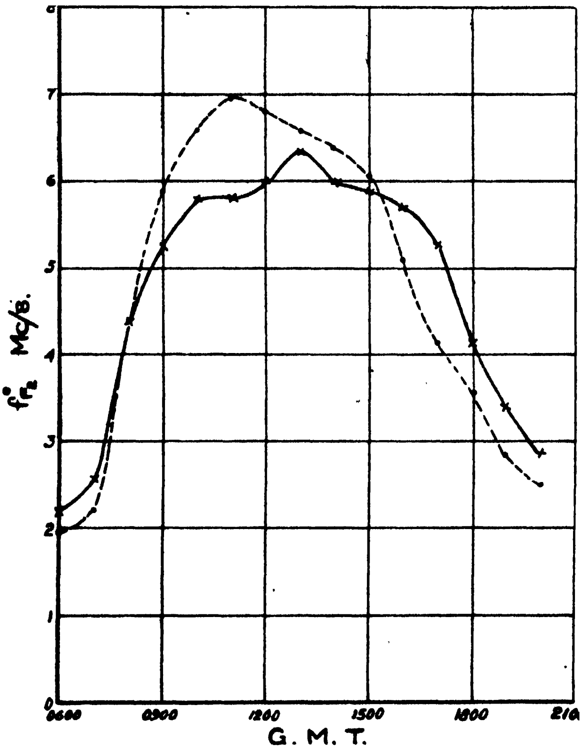


Figure 3. Mean values of $f^0_{F_2}$ at Slough.

30 Nov. 1942 to 12 Dec. 1942 •— — — —•

26 Jan. 1943 to 12 Feb. 1943 ×———×

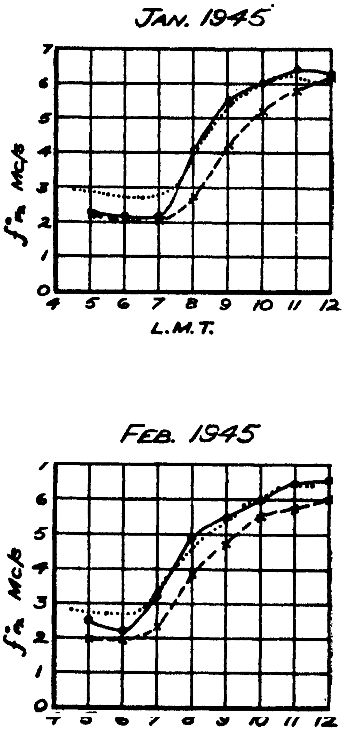


Figure 4. Monthly mean values of $f^0_{F_2}$.

×— — —× Burghead.

•·········• Moscow.

o———o Slough.

From many comparisons similar to those given above, it can be concluded that the time delay between corresponding ionosphere changes at points 500 km. apart, along a parallel of latitude at 51½° N., may be anything up to one hour or longer. The assumption of a fixed delay of 28 minutes, which has been made in the earlier part of this paper, is only valid when ionospheric conditions are changing quite rapidly, as is generally the case over the winter sunrise or sunset periods.

The experimental results for the Zeesen–Slough path may be summarized as follows.

Of the twenty-one values given in the first part of table 1, eight show an average positive error of 2·8% and thirteen an average negative error of 3·6%. It would appear that negative errors were slightly more probable than positive errors, but many of the errors are small and the relative numbers may not be significant. The arithmetic mean of these twenty-one calculated values differs from the observed value by −1·2% (probable error ±0·6%). This group of results suggests that for this distance and direction of propagation the systematic discrepancy between calculated and observed values of the M.U.F. does not exceed 2%. It will be noted that in nine cases the discrepancy between calculated and observed values exceeds the estimate of ±3% given in §2(c).

In the second group of observations there are only two cases in which the calculated M.U.F. exceeded that observed. These two positive errors have an average value of 2·8%, whereas the nine remaining negative errors have an average value of 9·3%. The arithmetic mean of the eleven calculated values is in error by −7·1%.

Considering all thirty-two observations together, we find that in ten cases the calculated M.U.F. exceeded that observed, the average discrepancy being 2·8%. In twenty-two cases the calculated M.U.F. was smaller than that observed, the average discrepancy being −6·0%. The average error (positive and negative) is about 5%. The arithmetic mean of the thirty-two calculated values is less than that observed by 3·2%.

(b) *Discussion of Moscow–Slough results*

Table 2 shows that in each period the average calculated value of the M.U.F. for the Moscow–Slough path, based on mean ionospheric data from Slough and Burghead, is some 11% too small.

We may note three possible causes for this discrepancy:

(i) For this distance of transmission the theoretical analysis may actually be in error by this amount.

(ii) At the critical sunrise period there might be appreciable lateral deviation of the incoming signal from the great-circle path.

(iii) The error may arise partly or entirely from incorrect assumptions about ionospheric conditions at the mid-point of the oblique path.

In considering these possibilities, it must be noted that an error of 11% is several times the magnitude of the error which we would have anticipated from the results for a distance of 1000 km., and would represent an extremely rapid rise in the divergence between theory and experiment when the distance is increased from 1000 km. to 2500 km. From the theoretical relationship between M.U.F. factor and distance for a "parabolic" reflecting layer, we should not expect the error at 2500 km. to be more than 1 or 2% larger than that observed for a distance of 1000 km.

The possibility of appreciable error due to abnormal lateral deviation of the signal at sunrise was investigated by taking bearing observations over the critical period on an Adcock direction finder. Tests on a sample of eight days, on each of which the calculated M.U.F. was much smaller than that observed, indicated conclusively that at the time of entry the ray signal was not deviated significantly from the true bearing value, and it would appear that in these particular experiments, this possibility is not a serious disturbing factor.

In considering the third possibility, we first note the results of a sample of observations on Moscow signals received at Burghead. The distance Moscow–Burghead (2460 km.) is only 2% less than the distance Moscow–Slough, so that the theoretical analysis should be equally accurate over the two paths. On the other hand the latitude of the mid-point of the Moscow–Burghead trajectory is practically equal to that of Burghead, and it should thus be possible to deduce ionospheric conditions at the mid-point from normal-incidence observations at Burghead alone. For measurements made at Burghead, there is thus reason to expect better agreement between theory and experiment than was noted in the case of the observations made at Slough. A group of ten such observations was made at Burghead in December 1943 and January 1944. In this case it was found that the mean calculated M.U.F. was 23% smaller than that observed. This very large error, under conditions in which we should have expected an error substantially smaller than 11%, suggests at once that in these experiments it is invalid to assume close correspondence between the local time variation and the longitude variation in ionosphere characteristics. It is also relevant to note that the error for these measurements at Burghead is almost exactly twice the magnitude of the error noted for the results given in table 2.

During the period of these Burghead observations a careful watch was also made at Slough on the time of entry of the Moscow ray signal. It was found that the mean time difference between the entry of the ray signal at Slough and Burghead was exactly equal to the time difference between sunrise at the two sites. Since the distances of Slough and Burghead from Moscow are practically equal, it could be inferred from this observation that, at the time of the experiments, ionosphere conditions governing the maximum usable frequencies for these two trajectories must have been identical.

The third point of interest is that for the 80 observations summarized in table 2, the mean M.U.F. calculated on Slough data alone is 10·63 Mc./s. (probable error ± 0·1 Mc./s.) and is thus within 3% of the correct value.

A little consideration of the three results given above will show that all three are quantitatively consistent with the conclusion that at the time of these experiments, ionospheric conditions governing the value of the M.U.F. at the mid-point of the Moscow–Slough trajectory, were very similar to those observed at Slough rather than to conditions midway between Slough and Burghead. Since these experiments were completed, substantial support for the correctness of this conclusion has come in the form of direct normal-incidence critical-frequency data from the ionospheric station recently in operation at Moscow. Figure 4 shows the diurnal variation in monthly mean values of $f^0_{F_2}$ measured at Slough, Burghead and Moscow for January and February 1945. Data from Moscow for December 1944 are, unfortunately, not available, but it is clear from the curves for January and February that over the period with which these experiments are concerned the monthly mean values of $f^0_{F_2}$ observed at Moscow are practically identical with the values observed at Slough.

Further confirmation of this type of variation in $f^0_{F_2}$ with longitude is provided when we examine critical-frequency data from other stations near this latitude but located further east. Thus if we consider data from Sverdlovsk (55° 50′ N., 60° 37′ E.) we find the winter values of $f^0_{F_2}$ at this station to be even higher than those observed at Moscow or at Slough. This marked longitude variation in critical frequency forms the subject of recent papers by Kessenikh and Bulatov (1944) and by Appleton (1946). It thus seems reasonable to assume that at the time of these experiments, near the mid-point of the Moscow–Slough path, conditions closely approximated to those given either at Moscow or Slough and, subject to this assumption, the mean observed value of the M.U.F. for 80 observations agrees with the mean calculated value to within 3%.

§ 6. ESTIMATION OF THE DISTANCE BETWEEN SENDER AND RECEIVER

In a paper by Appleton and Beynon (1942) a graphical representation has been given of the relation between M.U.F. factor and distance of transmission for a wide variety of ionospheric conditions. (The "M.U.F. factor" is the ratio of the M.U.F. to the normal-incidence critical frequency.) Hence if the M.U.F. factor be measured, it is a simple matter, from the curves, to estimate the distance between sender and receiver. At first it might appear that if the distance from the sender to receiver is unknown, then it will be impossible to measure the M.U.F. factor, since this implies a knowledge of ionospheric conditions at the mid-point of the trajectory. In the case of an east-west transmission path, however, we can overcome this difficulty by again assuming that the variation of ionosphere characteristics along the transmission path corresponds closely to the diurnal variation observed at vertical incidence at one end of the path. It is clear that these estimates of distance will be subject to similar errors as were the calculated values of the M.U.F. The actual estimates of sender-receiver distance for the Zeesen–Slough transmission path are given in the last column of table 1. For the whole group of thirty-two observations the mean calculated distance is 1036 km. The actual distance from Slough to Zeesen is 990 km., so that the actual error of the arithmetic mean is +46 km. or +4·6%. It is likely that if more accurate ionospheric data were available for the mid-point of the trajectory, then the accuracy of observations of sender-receiver distance made in this way would be considerably improved. If we consider the first group of results only, the mean calculated distance is 1005 km., so that the actual error in this case is only 15 km. or +1·5%.

It may be noted that for ranges beyond about 1000 km., and for a given accuracy in estimating the maximum usable frequency factor, there is theoretically a steady deterioration with increasing distance in the accuracy of a single estimate of sender-receiver distance. Thus an error of 2% in the M.U.F. factor for the Moscow–Slough path corresponds to a distance error of about 100 km.

§ 7. CONCLUSIONS

One of the objects of these experiments was to investigate the magnitude of the errors to be expected in a practical application of maximum usable frequency calculations. In the present experiments, full normal-incidence ionospheric data were available for one end of the transmission path and the transmitters were located in a direction approximately east of the receiving site. The conditions were thus rather more favourable than can be expected in the general application of such calculations. Nevertheless the results indicate that even for a transmission path of 1000 km., and with full normal-incidence ionospheric data for one end of the path, the divergence between individual calculated and observed values may occasionally amount to 15%. Results for the longer path of 2500 km. show differences of up to 25% between calculated and observed values in individual cases. In the case of the experiments over 1000 km., the fact that positive and negative errors were equally frequent suggests that there is no serious error in the analysis underlying the M.U.F. calculation. In the case of the measurements over 2500 km., the mean of 80 calculated values was 11% less than the observed value. The calculated value exceeded the observed value in seven cases only. There is evidence, however, that this discrepancy is almost entirely due to incorrect assumptions about ionospheric conditions near the mid-point of the trajectory. If such data are available it seems likely that the discrepancy between calculated and observed values of the maximum usable frequency for a distance of 2500 km. would not exceed 3%.

§ 8. ACKNOWLEDGMENTS

The results for the Zeesen transmissions described above were originally included in a confidential paper communicated in April 1943 to the Radio Research Board. The work was carried out as part of the programme of the Radio Research Board, and this paper is published by permission of the Department of Scientific and Industrial Research.

REFERENCES

Appleton, E. V., 1946. *Nature, Lond.*, **157**, 691.
Appleton, E. V. and Beynon, W. G., 1940. *Proc. Phys. Soc.*, **52**, 518.
Appleton, E. V. and Beynon, W. G. Radio Research Board, June 1942. (In course of publication.)
Breit, G. and Tuve, M. A., 1926. *Phys. Rev.*, II, **28**, 554.
Kessenikh, V. N. and Bulatov, H. D., 1944. *C.R. (Doklady) Acad. Sci. U.R.S.S.*, **45**, No. 6, 234.
Martyn, D. F., 1935. *Proc. Phys. Soc.*, **47**, 323.
Millington, G., 1938. *Proc. Phys. Soc.*, **50**, 801.
Schafer, J. P. and Goodall, W. M., 1939. *Terr. Mag. Atmos. Elect.*, **44**, 205.
Smith, N., 1938. *J. Res. Nat. Bur. Stds., Wash.*, **20**, 683.

DISCUSSION

on papers by:

(i) Sir E. Appleton and W. J. G. Beynon, F.R.S.: "The application of ionospheric data to radio communication problems" (*Proc. Phys. Soc.*, **59**, 58 (1947)).

(ii) W. J. G. Beynon: "Oblique radio transmission in the ionosphere and the Lorentz polarization term" (*Proc. Phys. Soc.*, **59**, 97 (1947)).

(iii) W. J. G. Beynon: "Some observations of the maximum frequency of radio communication over distances of 1000 km. and 2500 km." (*Proc. Phys. Soc.* **59**, 521 (1947)).

Mr. R. Dehn. With reference to the agreement of observations of overhead conditions of the ionosphere with those obtained from records of signals received and reflected at a distance equivalent to some 25 or 30 min. in time, would these conditions of the ionosphere be preserved for a longer period of time? That is, would the agreement be observable over greater distances?

Mr. F. A. Kitchen. Could prediction of maximum usable frequencies be made as much as twelve months ahead, following the correlation of distance factors with sunspot variation?

Mr. R. Naismith. A forecast of maximum usable frequencies for a period of twelve months ahead involves two quite separate requirements: an estimate of the vertical incidence conditions twelve months ahead; and an estimate of the maximum usable frequency from given vertical incidence conditions.

The authors have succeeded in supplying a very satisfactory solution to the latter requirement and have shown that the resultant error can be less than 3 %.

An illustration of the comparative accuracy of the former may be given from current data. In any estimate of the vertical incidence critical frequency, account must be taken of the variation in the solar cycle among other factors. It is well known that the ionization in region F_2 varies in sympathy with the solar cycle when average values are coinsidered It may be assumed, therefore, that the ionization over the world would vary fairly uniformly due to this cause. Noon values of ionization for region F_2 in different parts of the world were compared for September 1945 and September 1946 (the latest month for which data are available), and it was found that whereas a 10 % increase occurred in one part of the world, a 90 % increase occurred in another. On present knowledge this large variation was quite unpredictable, and illustrates one of the difficulties in producing a forecast of ionospheric conditions so far ahead.

Dr. F. T. Farmer. I would like to know whether the calculations described take into account the earth's magnetic field for the ordinary wave, or whether it is only allowed for in calculating oblique frequencies for the extraordinary wave.

Author's reply. In reply to Mr. Dehn, it is possible that certain forms of ionospheric variation would persist for longer periods. Thus the detailed fluctuation in the sunrise change shown in figure 1 (p. 529) might occur at widely separated stations. Clear evidence of such correlation was noted only in the Zeesen–Slough experiments. In the case of the Moscow–Slough observations the repetition of detailed ionospheric changes was not obvious, but it is possible that a careful examination of the records would yield further information on this matter.

Concerning Mr. Kitchen's point, I think that Mr. Naismith has largely provided the answer. The main purpose of these papers was not to discuss prediction, but I may say that some predictions of the maximum usable frequency have been made for six months ahead, and, subject to a proportionate decrease in accuracy, an extension to a period of twelve months could, no doubt, be made. However, Mr. Naismith has already pointed out some of the difficulties involved.

In reply to Dr. Farmer, the effect of the magnetic field of the earth was not included in the calculations relating to the ordinary wave.

MEAN FREE PATH OF SOUND IN AN AUDITORIUM

By A. E. Bate and M. E. Pillow,
Northern Polytechnic, London

MS. received 11 *October* 1946

ABSTRACT. A brief review of the subject is followed by proofs that the mean free path of sound in an enclosure is equal to 4(Volume/Surface Area) for rectangular, spherical, and cylindrical rooms of any dimensions.

§ 1. INTRODUCTION

In some methods (Eyring, 1930) of calculating the reverberation time of sound in an auditorium in terms of the dimensions of the room and the absorptive properties of the walls, it is necessary to make use of a formula for the mean free path of sound; i.e. the mean distance travelled by all "rays" of sound between successive impacts with the walls.

Jaeger (1911), applying a method similar to that used in the kinetic theory of gases, obtained for this mean free path the value $4V/S$, where V represents the volume and S the internal surface area of the room. Jaeger's method shows that the value should be independent of the shape of the room and the position of the source, if uniform distribution of the sound energy is assumed: that is, if sound is travelling with equal intensity in all directions, through all points in the enclosure, at any instant considered.

Schuster and Waetzmann (1929) calculated the mean free path for rooms of certain simple shapes. Their values, which Eyring (1930) obtained at almost the same time by more approximate methods, are:

Cube	M.F.P. $= 2\sqrt{3}\,V/S = 3\cdot5\,V/S$,
Cylinder (height = diameter)	M.F.P. $= 3\sqrt{2}\,V/S = 4\cdot2\,V/S$,
Sphere	M.F.P. = diameter $= 6\,V/S$.

These values, however, were calculated by choosing arbitrarily the position of the source, or the direction of emission of the sound, so the energy distribution would not be uniform.

Knudsen (1932) carried out experiments designed to show that for rooms of the usual shapes the mean free path of sound is independent of the shape. With the help of a flash-lamp and mirror, he used light rays in place of sound, in scale models of the auditoria considered, and traced the paths followed by successively reflected rays emitted in a limited number of evenly distributed directions, and averaged the distance travelled between reflections, taking the same number of reflections for each emitted ray. His results agree to a good approximation in most cases with Jaeger's formula $4V/S$.

It is the purpose of this paper to show, by direct averaging, that the mean free path of sound in (*a*) *any* rectangular enclosure, (*b*) a spherical enclosure, (*c*) *any* cylindrical enclosure, is $4V/S$.

§ 2. OUTLINE OF METHOD

When a uniform source of sound has been active in an enclosure for an appreciable time, the sound energy density may be assumed uniform throughout, i.e. the energy is travelling uniformly in all directions from all points in the enclosure (with the exception of certain special cases in which "focusing" occurs). Again, the coefficient of absorption is assumed to be sufficiently low for many reflections to take place, and the sound disturbance between reflections is assumed to travel in straight lines with constant speed until its energy is dissipated.

Now the sound energy may be considered to consist of very small units, or quanta, which move in straight lines with speed c between impacts with the walls, and which do not interfere with each other's motion in any way. (It is not suggested that these quanta are indivisible units.)

Suppose that, in a very small interval of time after the energy density has become uniform in the sense indicated above, $4\pi n$ "quanta" (i.e. n per unit solid angle) leave each of a large number of points which are uniformly distributed throughout the enclosure with volume density ρ.

The method consists in finding the total number of impacts with the interior surfaces of the room made by all these quanta during a finite time T immediately following that short interval, and dividing the total distance $4\pi n\rho cT$, covered by all of the quanta, by this number of impacts. For convenience, the number of impacts, and the distance covered, *per second*, have been used in the following proofs.

§ 3. RECTANGULAR ENCLOSURE

Dimensions of room are a, b, h, with axes of coordinates parallel to edges of room as shown.

Speed of each quantum $= c$.

A quantum projected from any point S in the direction (θ, ϕ) has velocity components $c \sin\theta \cos\phi$, $c \sin\theta \sin\phi$, $c \cos\theta$, in the x, y, z directions.

In the x-direction the distance travelled between impacts with walls is a.

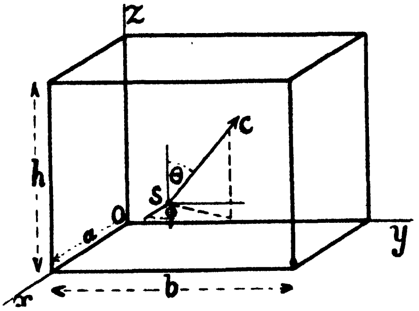


Figure 1.

$\therefore$ Number of impacts per sec. with walls perpendicular to the x-axis is

$$(c/a)\sin\theta\cos\phi.$$

Similarly, numbers of impacts per second with y-walls and z-walls are

$$(c/b)\sin\theta\sin\phi \quad \text{and} \quad (c/h)\cos\theta.$$

$\therefore$ Total number of impacts made per second by one quantum

$$=(c/a)\sin\theta\cos\phi+(c/b)\sin\theta\sin\phi+(c/h)\cos\theta.$$

Now the elementary solid angle with its axis in the (θ, ϕ) direction is

$$\sin\theta \,.\, d\theta \,.\, d\phi,$$

and the number of quanta projected from one point in the (θ, ϕ) direction may therefore be taken as

$$n \sin\theta \,.\, d\theta \,.\, d\phi.$$

$\therefore$ Total number of impacts per second made by quanta from one point is

$$8n \int_0^{\pi/2} \int_0^{\pi/2} \left\{ \frac{c}{a} \sin\theta\cos\phi + \frac{c}{b}\sin\theta\sin\phi + \frac{c}{h}\cos\theta \right\} \sin\theta \,.\, d\theta \,.\, d\phi,$$

and by direct integration this becomes

$$2n\pi\,[c/a+c/b+c/h]$$

$$=2n\pi c \,.\, (bh+ha+ab)/abh.$$

But area of walls $=2(bh+ha+ab)=S$ and volume of room $=abh=V$.

$\therefore$ Number of impacts per second made by quanta from one point

$$=\pi nc \,.\, S/V.$$

The total number of quanta projected from this point is $4\pi n$, and each travels a distance c per second.

$\therefore$ Total distance covered per second by these quanta is $4\pi nc$.

$\therefore$ Mean distance travelled between impacts

$$=4\pi nc \div \pi nc\,.\,S/V=4\,.\,V/S.$$

This is independent of the position of the point of projection, and is, therefore, the same for all such points.

$\therefore$ Mean free path $=4V/S$.

§ 4. SPHERICAL ENCLOSURE

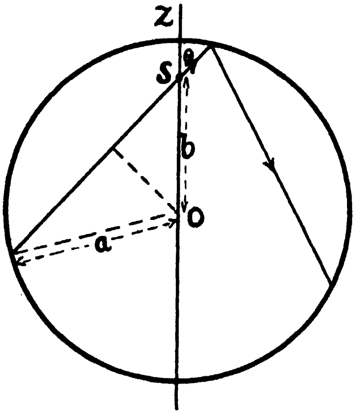


Figure 2.

Radius of sphere $=a$.

z-axis is a diameter, and O the centre.

Consider first a quantum projected with speed c in (θ,ϕ) direction from a point S on the z-axis at distance b from the centre. After reflection, this quantum will describe a number of equal chords, each of length

$$2\sqrt{a^2-b^2\sin^2\theta}.$$

$\therefore$ Number of impacts made per second by this quantum

$$=c/(2\sqrt{a^2-b^2\sin^2\theta}).$$

As before, $n\sin\theta\,.\,d\theta\,.\,d\phi$ quanta are projected into the elementary solid angle whose axis is in the (θ,ϕ) direction.

$\therefore$ Total number of impacts made per second by all the $4\pi n$ quanta projected from S

$$=\int_0^\pi\int_0^{2\pi}\frac{nc\sin\theta\,.\,d\theta\,.\,d\phi}{2\sqrt{a^2-b^2\sin^2\theta}}$$

$$=\pi nc\int_0^\pi\frac{\sin\theta\,.\,d\theta}{\sqrt{(a^2-b^2)+b^2\cos^2\theta}}$$

$$=\frac{\pi nc}{b}\left[-\sinh^{-1}\left(\frac{b}{\sqrt{a^2-b^2}}\cos\theta\right)\right]_0^\pi$$

$$=\frac{2\pi nc}{b}\,.\,\sinh^{-1}\frac{b}{\sqrt{a^2-b^2}}$$

$$=\frac{\pi nc}{b}\log_e\frac{a+b}{a-b}.$$

This expression represents the number of impacts made per second by $4\pi n$ quanta.

The total distance covered by them per second is $4\pi nc$.

$\therefore$ Mean distance travelled batween impacts by all quanta projected from S is

$$4b/\log_e\frac{a+b}{a-b}.$$

By substituting a series of values for b, it is seen that the value of this mean free path varies with the point of projection, having values ranging between 0 and $2a$.

If $b \to 0$, the mean free path becomes

$$\operatorname*{Lt}_{b\to 0} \frac{4b}{\dfrac{2b}{a}+\dfrac{2}{3}\dfrac{b^3}{a^3}+\ldots}$$

i.e. $$2a \text{ or } 6V/S.$$

This agrees with the result obtained by Schuster and Waetzmann for a sphere, and it can be seen that if the source of sound is located at the centre of the sphere all the "rays" will be reflected back to this point, and the uniform distribution of energy will never be attained. This is a special case of "focusing", and a formula obtained by placing the source at the centre of the sphere cannot be applied to the general case.

If the source is at any point other than the centre, the energy distribution will be uniform when the source has been sounding long enough for a considerable number of reflections to have occurred—the condition assumed throughout—and the points of projection of the "quanta", as explained above, will then be distributed over the whole volume.

Suppose that there are ρ such points of projection per unit volume. Then the number lying at a distance between b and $b+db$ from O is $\rho \,.\, 4\pi b^2 \,.\, db$.

The number of impacts made per second by quanta projected from one such point has been shown to be

$$\frac{\pi nc}{b} \,.\, \log_e \frac{a+b}{a-b}.$$

$\therefore$ Number of impacts made per second by quanta projected from all points in the enclosure is

$$\int_0^a \frac{\pi nc}{b} \,.\, \log_e \frac{a+b}{a-b} \,.\, \rho \,.\, 4\pi b^2 \,.\, db,$$

which reduces to $4\pi^2 \rho nca^2 = \pi\rho nc \,.\, S$, where S is the surface area.

But the total number of quanta is $\rho V \,.\, 4\pi n$, so the total distance covered per second is $4\pi\rho nc \,.\, V$.

$\therefore$ Mean free path $= 4\pi\rho nc \,.\, V \div \pi\rho nc \,.\, S = 4V/S$.

§ 5. CYLINDRICAL ENCLOSURE

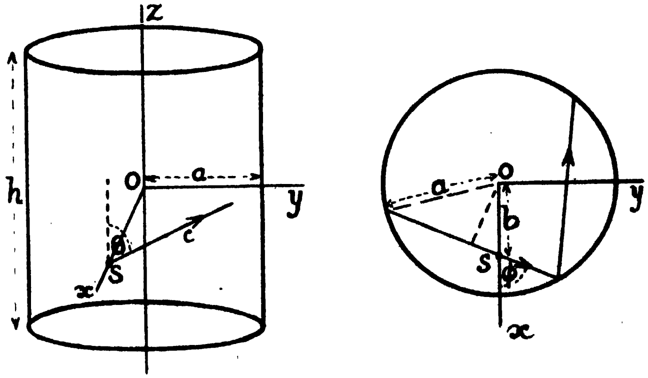


Figure 3.

Cylinder of height h and radius a. Axis of cylinder is z-axis.

Consider a quantum projected with speed c in the (θ, ϕ) direction from the point S, at a distance b from the x-axis. For convenience in reference, let S lie on the x-axis.

Component velocity parallel to the z-axis $= c\cos\theta$.

Distance traversed between impacts in this direction $= h$.

∴ Number of impacts made per second with ends of the cylinder

$$= (c/h)\cos\theta.$$

Component velocity perpendicular to the z-axis, i.e. in x-y plane,

$$= c\sin\theta.$$

Projection of path on this plane consists of a series of equal chords, each of length

$$2\sqrt{a^2 - b^2\sin^2\phi}.$$

∴ Number of impacts made per second with curved walls

$$= \frac{c\sin\theta}{2\sqrt{a^2 - b^2\sin^2\phi}}.$$

∴ Total number of impacts made per second by one quantum

$$= \frac{c}{h}\cos\theta + \frac{c\sin\theta}{2\sqrt{a^2 - b^2\sin^2\phi}}.$$

As before, $n\sin\theta \,.\, d\theta \,.\, d\phi$ quanta are projected into the elementary solid angle with its axis in the (θ, ϕ) direction.

∴ Total number of impacts made per second by quanta projected from S

$$= n\int_0^\pi \int_0^{2\pi} \left\{\frac{c\cos\theta}{h} + \frac{c\sin\theta}{2\sqrt{a^2 - b^2\sin^2\phi}}\right\} \sin\theta \,.\, d\theta \,.\, d\phi$$

$$= 8cn\int_0^{\pi/2} \int_0^{\pi/2} \left\{\frac{\sin 2\theta}{2h} + \frac{1 - \cos 2\theta}{4\sqrt{a^2 - b^2\sin^2\phi}}\right\} d\theta \,.\, d\phi$$

$$= \frac{2\pi cn}{h} + \pi cn\int_0^{\pi/2} \frac{d\phi}{\sqrt{a^2 - b^2\sin^2\phi}}.$$

This expression represents the number of impacts made per second by the $4\pi n$ quanta projected from S.

Suppose, as before, that the points of projection are uniformly distributed with volume density ρ.

∴ The number of such points at distances from the axis of the cylinder between b and $b + db$ is

$$\rho \,.\, 2\pi h \,.\, b \,.\, db.$$

∴ Total number of impacts made per second by quanta projected from all points in enclosure

$$= \int_0^a \left\{\frac{2\pi cn}{h} + \pi cn\int_0^{\pi/2} \frac{d\phi}{\sqrt{a^2 - b^2\sin^2\phi}}\right\} \rho \,.\, 2\pi h \,.\, b \,.\, db,$$

$$= 2\pi^2\rho cna^2 + 2\pi^2\rho cnh\int_0^{\pi/2} \int_0^a \frac{d\phi \,.\, b\,db}{\sqrt{a^2 - b^2\sin^2\phi}},$$

which on integration becomes

$$\pi\rho cn[2\pi a^2 + 2\pi ha] = \pi\rho cn \,.\, S,$$

where S is the surface area.

This is the total number of impacts made per second.

But total number of quanta is $\rho V . 4\pi n$, so that the total distance travelled per second is

$$\rho V . 4\pi nc.$$

$\therefore$ Mean free path $= 4\pi\rho nc . V \div \pi\rho nc . S = 4V/S$.

REFERENCES

EYRING, 1930. *J. Acoust. Soc. Amer.*, **1**, 217.
JAEGER, 1911. *Wiener Akad. Ber. Math. Naturw. Kl.*, **120**, II *a*.
KNUDSEN, 1932. *Architectural Acoustics* (London : John Wiley), Chapter V.
SCHUSTER and WAETZMANN, 1929. *Ann. Phys., Lpz.*, **1**, 5, 671.

DETERMINATION OF THE CRYSTAL STRUCTURE OF GOLD LEAF BY ELECTRON DIFFRACTION

BY T. B. RYMER AND C. C. BUTLER*,
The University, Reading
* Now at the University of Manchester

MS. received 10 *April* 1946 ; *in revised form* 15 *January* 1947

ABSTRACT. It is found that the radii of the rings of Debye-Scherrer electron-diffraction photographs obtained from gold leaf are not in exact agreement with the theoretical values. This is ascribed to the crystal lattice being distorted by surface-tension forces.

§ 1. INTRODUCTION

THE radii of the Debye-Scherrer rings of electron-diffraction patterns are generally assumed to agree exactly with the predictions of simple theory. So far as we are aware, no attempt has been made to make measurements of higher accuracy than one or two parts in a thousand. Usually, the rings have a radius of the order of one or two cm. and are measured with a travelling microscope having 10-micron divisions ; most observers are satisfied if their readings agree to within a few divisions. There are two main reasons for failure to obtain a higher accuracy. First, plates with a finer grain and microscopes with higher magnification than usual are required. Secondly, if attempts are made to test precision by measuring the radius of a single ring in different azimuths or by comparing the relative radii of different rings, discrepancies are observed which appear to indicate measurement errors of the order of a few parts in a thousand. However, as a result of extensive measurements and stringent statistical tests, we have established that the present practical limit of precision is really of the order of one part in ten thousand, and that the apparent discrepancies alluded to are due to features of the diffraction pattern which are not dealt with by present theories. These peculiarities are of two kinds: the rings are not exactly circular and their radii are not given exactly by Bragg's law.

In a recent paper (Rymer and Butler, 1945 a), we have given a general discussion of the problem of measuring ring radii and have presented some evidence in support of our claim to be able to make measurements of high accuracy. Some of the purely instrumental effects responsible for the abnormal radii and non-circular form of rings are discussed in another paper (Rymer and Butler, 1945 b); the analysis of these features provides an independent check on the precision of measurement.

Our present purpose is to discuss a peculiarity of the diffraction pattern of gold leaf which is revealed by these high-precision measurements: that the relative radii of the rings differ from the expected values by amounts of the order of 0·05 per cent. In view of the crucial importance of an accurate estimate of the precision of measurement, we give a brief description of the experimental technique and we present additional evidence showing the accuracy attained.

§ 2. EXPERIMENTAL

The diffraction camera used was of the type described by Finch, Quarrell and Wilman (1935) with minor modifications. The high-tension supply was a half-wave rectifier set with the output fed through a saturated diode. An additional 500-pF. condenser connected across the camera discharge tube served to reduce voltage fluctuations. Oscillographic examination showed that the ripple voltage was 75 volts r.m.s. at 100 c.p.s. when the discharge tube was taking its normal load of $\frac{1}{2}$ ma. at 50 kv.

Particular care was taken to eliminate stray alternating magnetic fields by the use of compensating coils carrying currents of suitable magnitude and phase. Tests with a search coil and oscillograph showed that the alternating magnetic field at the axis of the camera nowhere exceeded a peak value of $10\,\gamma$. It has been shown (Rymer and Butler, 1945 b) that the slight broadening of the rings introduced by the combined action of alternating fields and high-tension ripple is incapable of introducing errors into the radius measurements of more than $\frac{1}{2}\,\mu$.

Ilford Thin Film Half Tone plates were used with a borax fine-grain developer. The uniformity of this plate and the fineness of the grain combine to make it superior for this purpose to any sofar tried. Plates with coarser grain, such as Ilford Ordinary or Special Rapid, are quite unsuitable for precision work. Tests show that with a constant intensity of electrons of about 50 kv. energy, the density produced on a Thin Film plate is accurately proportional to the exposure time up to densities of at least unity; since the reciprocity law holds for electrons (Becker and Kipphan, 1931), the density is proportional to the intensity. This is a matter of importance, since otherwise systematic errors can occur in the measurement of the ring radius (Rymer and Butler, 1945 a).

The width of the beam where it strikes the photographic plate has been measured (Rymer and Butler, 1945 b) and found to be $40\,\mu$. The width of the diffraction rings produced by gold leaf is of the order of $200\,\mu$, and is therefore due almost entirely to broadening by the crystal; there is therefore little to be gained by further attempts to sharpen the electron beam. That the broadening is due to the finite size of the crystals rather than to a variable lattice constant is suggested by the data of table 1, which show that the ring width is independent of the radius.

Table 1. Plate C/279. Gold-leaf specimen. Widths of diffraction rings

Indices	111	200	220	311
Width (microns)	140	116	110	130

Measurements of the ring radii were made with an instrument reading to one micron (Rymer and Butler, 1944) which can be used as a travelling microscope or as a non-recording microphotometer and is fitted with a divided head which permits azimuths of radii measurements to be determined to 0°·5. Corrections have to be applied to the measured radii to allow for the following effects:

(*a*) Errors of microphotometer screw.

(*b*) Background density due to incoherent scattering of electrons.

(*c*) Curvature effect: the intensity is enhanced on the inner side of a diffraction ring owing to its shorter perimeter.

(*d*) Finite length of microphotometer slit.

The method of determining these corrections has previously been described (Rymer and Butler, 1945 a).

§ 3. PRECISION OF MEASUREMENT

Before presenting the results of measurements on gold specimens, it is desirable to discuss the evidence of the accuracy of the technique provided by substances which do not exhibit any peculiarities. Our belief in the reliability of our measurements is based on the agreement between the values of the standard deviation as computed by the following completely independent methods:

(*a*) Analysis of the scatter of repetition readings.

(*b*) Analysis of the scatter of measurements of the radius of a single ring in different azimuths when allowance is made for instrumental effects.

(*c*) Comparison of the variation of the ellipticity of the diffraction rings due to stray magnetic fields with theoretical predictions.

(*d*) Comparison of the mean radii of diffraction rings with theoretical values.

This is illustrated by measurements made on Plate No. D/137 of a sodium chloride specimen (figure 4).

(*a*) The plate was measured with the microscope by making three settings on each ring and repeating the process in each of eighteen equally spaced azimuths. From the scatter of each group of three settings from their mean, the standard deviation of a single setting on the (111), (200), (220) and (420) rings is found to be 2·6, 2·2, 2·2 and 2·9 μ respectively. The average standard deviation of a single setting is thus

$$\text{S.D.} = 2{\cdot}49 \pm 0{\cdot}15\,\mu.^{*} \qquad \ldots\ldots(1)$$

(*b*) The radius of the (200) ring was measured in azimuths 0, 20, 40, ... 340 degrees using the microscope, the mean of three readings being taken in each case. Table 2 shows the results, the radii being given in microns.

Table 2. Radius of (200) ring = 8790 + following

Azimuth	0	20	40	60	80	100	120	140	160
Radius	33	26	21	18	9	5	10	22	25
Azimuth	180	200	220	240	260	280	300	320	340
Radius	37	41	53	60	65	68	51	50	39

---

* Throughout this paper, ± indicates *standard deviation*.

Fitting a Fourier series to these results gives *

$$\text{Radius} = 8825{\cdot}20 - 26{\cdot}40 \sin(\theta + 1{\cdot}78)^\circ + 2{\cdot}49 \sin(2\theta + 327{\cdot}79)^\circ. \qquad \ldots\ldots(2)$$

The first harmonic is trivial, being due to measurements being made from a point which is not quite the centre of the pattern. The second harmonic is due to the presence of stray magnetic fields (Rymer and Butler, 1945 b). From the differences between the radii given by equation (2) and the numbers in table 2, we obtain the standard deviation of any radius measurement as $3{\cdot}75 \pm 0{\cdot}73\,\mu$. Since each radius measurement is the mean of three readings, the standard deviation of a single reading must be $\sqrt{3}$ times this:

$$\text{S.D.} = 6{\cdot}49 \pm 1{\cdot}27\,\mu. \qquad \ldots\ldots(3)$$

(*c*) Theoretically, the amplitude of the second harmonic in the expression for the radius should be proportional to the mean radius, while the phase should be the same for all rings (Rymer and Butler, 1945 b). Table 3 gives the amplitude and phase of the harmonic for four rings.

Table 3

Indices	Amplitude (microns)	Phase (°)	Mean radius (microns)	$\frac{\text{Amplitude}}{\text{Radius}}$
111	3·2	244·1	7645·6	$4{\cdot}19 \times 10^{-4}$
200	2·5	327·8	8827·1	2·83
220	7·0	290·0	12477·0	5·61
420	8·7	280·3	19731·3	4·41

The larger harmonics naturally give more accurate values for the amplitude/radius. We therefore weight the results in proportion to the radius, obtaining for the mean amplitude/radius $4{\cdot}40 \times 10^{-4}$. From the differences between this and the numbers in the fifth column of table 3 it can be deduced that the standard deviation of a single harmonic is $1{\cdot}18 \pm 0{\cdot}48\,\mu$. Now it can be shown that the standard deviation of the amplitude of a harmonic computed from N radii measurements is $\sqrt{(2/N)}$ times the standard deviation of a single radius measurement. Since there are 18 radii measurements, and each is the mean of three readings, we obtain for the standard deviation of a single reading

$$\text{S.D.} = 6{\cdot}15 \pm 2{\cdot}51\,\mu. \qquad \ldots\ldots(4)$$

An estimate of the precision can also be made from the phase angles of the second harmonic. The weighted mean phase is 285°·72. Now the standard deviation of the phase (measured in radians) is equal to the standard deviation of the amplitude of the harmonic divided by the amplitude. Using this relation, it can be calculated from the data of table 3 that the standard deviation of the amplitude of a single harmonic is $2{\cdot}23 \pm 0{\cdot}91\,\mu$, corresponding to a standard deviation of a single reading of

$$\text{S.D.} = 11{\cdot}60 \pm 4{\cdot}73\,\mu. \qquad \ldots\ldots(5)$$

---

* It must be emphasized that the fact that the coefficients in this series are given to the second decimal place does *not* imply that their value is necessarily known even to the first decimal place. The coefficients have been calculated to the second decimal place merely in order to avoid any possibility of "rounding-off" errors in the determination of the standard deviation. Similar remarks apply to *all figures quoted in this paper*. The *only* valid estimate of accuracy is based on a calculation of the standard deviation.

(*d*) If d is the interplanar spacing, R the radius of the diffraction ring, L the distance from specimen to photographic plate and λ the wave-length of the electrons, then theoretically

$$Rd=\lambda L+\tfrac{3}{8}\frac{\lambda L}{L^2}R^2. \qquad \ldots\ldots(6)$$

For several reasons this relation does not hold exactly, but the values of λL calculated from this equation vary systematically with the ring radius according to the relation (Rymer and Butler, 1945 a and b)

$$\lambda L=(\lambda L)_0+B/R^2, \qquad \ldots\ldots(7)$$

where B is a constant. In our first paper (1945 a) we have shown that the experimental results for the plate under discussion fit these equations, the standard deviation of a single λL determination being $2{\cdot}2\times10^{-12}$ cm², while $(\lambda L)_0=2{\cdot}48093\times10^{-8}$ cm². Since the average radius R is 1·2170 cm., this gives for the standard deviation of a mean radius 1·09 μ. A rather more satisfactory procedure is to weight the (λL) determinations in proportion to the radius. When this is done, and the constants of equation (7) are determined by the method of least squares, it is found that the standard deviation of the mean radius of a ring is $0{\cdot}87\pm0{\cdot}43\,\mu$. This is made up of two parts: random errors in the individual radii measurements and errors in the determination of the systematic corrections. If the latter are neglected, we obtain for the standard deviation of a single reading

$$\text{S.D.}=6{\cdot}39\pm3{\cdot}19\,\mu. \qquad \ldots\ldots(8)$$

This is almost certainly too high as the error in the systematic corrections is unlikely to be entirely negligible. A reasonable estimate (Rymer and Butler, 1945 a) of this latter error is 0·77 μ. Using this value, the standard deviation of a single reading becomes

$$\text{S.D.}=2{\cdot}99\pm1{\cdot}49\,\mu. \qquad \ldots\ldots(9)$$

The higher value of the standard deviation derived from the study of the amplitude and phase of the second harmonic (equations (4) and (5)) may be due to the stray magnetic fields not being *linear* functions of position as is assumed in the theory which predicts that the amplitude should be proportional to the radius and that the phase should be the same for all rings. Unpublished results from a large number of photographs of many substances indicate that in general the diffraction rings are not circular and that the expression for the radius given by equation (2) should be extended by the addition of higher harmonics whose origin will be discussed in a later paper. Owing to the presence of these harmonics, the estimate of the standard deviation given in equation (3) is too high. Comparing the results for the standard deviation of a single reading given by equations (1), (3), (4), (5) and (9), we see that this quantity is certainly less than 6 μ and that a very fair estimate is 3 μ.

Naturally, few plates have beeen studied so exhaustively as the one discussed. Less complete measurements have been made on many plates for the study of special features, and these all confirm the estimate given of the precision of measurement for rings of this width. Since the width of the rings and general quality of the plate under discussion is very little different from that of the gold-leaf diffraction patterns to be discussed (compare figure 4 with figures 5 and 6), we may take 3 μ as a reasonable estimate of the standard deviation of a single reading for the latter plates.

§4. GOLD-LEAF DIFFRACTION PATTERNS

When an attempt is made to fit equations (6) and (7) to measurements of gold-leaf diffraction patterns, discrepancies are immediately observed. The nature of these can be illustrated by the results for a typical plate (figure 5) of a diffraction pattern of a specimen of gold leaf which had been thinned in potassium cyanide solution. Table 4 gives the mean radius R of the first four diffraction rings and the values of λL calculated from them by means of equation (6), the lattice constant of gold being taken as 4·0700 A. and the camera length L as 47 cm. Figure 1 is a graph of λL against $1/R^2$. The dotted line has been fitted by the method of least squares, the different points being given weights proportional to R, and has the equation (R in cm.)

$$\lambda L = (22836{\cdot}47 + 13{\cdot}77/R^2) \times 10^{-12}\,\text{cm}^2 \qquad \ldots\ldots(10)$$

The values of λL calculated from this are listed in table 4 under $(\lambda L)_{\text{cal.}}$ (eqn. (10)),

Table 4. Plate No. D/268

Indices	R (microns)	λL 10^{-12} cm^2	$(\lambda L)_{\text{calc.}}$ (eqn. (10))	$(\lambda L)_{\text{calc.}}$ (eqn. (11))
111	9724·8	22847·8	22851·03	22847·67
200	11232·4	53·0	47·38	44·94
220	15879·0	39·5	41·93	40·84
311	18623·9	40·8	40·44	39·62

and from the differences between these and the observed λL it can be calculated that the standard deviation of each radius determination is $2{\cdot}50 \pm 1{\cdot}25\,\mu$. Of this amount, $0{\cdot}77\,\mu$ arises from the error in the systematic corrections (Rymer and Butler, 1945 a). Also, a radius determination is the mean of 54 readings, for the plate was measured along 18 equally spaced azimuths and in each case three settings were made on the rings. Hence the standard deviation of a single setting must be $\sqrt{54}\,\sqrt{2{\cdot}50^2 - 0{\cdot}77^2} = 17{\cdot}5 \pm 8{\cdot}8\,\mu$. This is distinctly greater than the expected value of $3\,\mu$.

Examination of figure 1 suggests that the points corresponding to the (111), (220) and (311) rings lie much more nearly on a straight line than do all four points. The full line has been fitted to these points by the method of least squares and has the equation

$$\lambda L = (22836{\cdot}73 + 10{\cdot}35/R^2) \times 10^{-12}\,\text{cm}^2 \qquad \ldots\ldots(11)$$

The values of λL calculated from this are listed in table 4 under $(\lambda L)_{\text{calc.}}$ (eqn. 11), and from the differences between them and the observed λL it can be calculated that the standard deviation of a radius determination (allowing for errors of the systematic corrections) is $1{\cdot}03\,\mu$, corresponding to a standard deviation of a single setting of $7{\cdot}5 \pm 5{\cdot}3\,\mu$. This is much closer to the expected value. On the other hand, equation (11) implies that the observed value of λL for the (200) diffraction is $(8{\cdot}07 \pm 2{\cdot}76) \times 10^{-12}$ cm^2 larger than would be expected.

We have so far measured over a dozen diffraction patterns of gold leaf prepared under various conditions, and in *every* case the value of λL calculated from the (200) diffraction ring is higher than would be expected from the results for the remaining rings. In table 5 the magnitude of this anomaly is given in the third column. The form of the variation of the anomaly with the wave-length of the electrons is

Table 5. (200) diffraction anomaly

Plate No.	$(\lambda L)_0$ (10^{-12} cm^2)	Anomaly (10^{-12} cm^2)	$\frac{\text{Anomaly}}{(\lambda L)_0}$ ($\times 10^{-4}$)	Remarks
D/257	22510·32	8·73	3·88	1938 gold. Amalgam rings
D/268	22836·73	8·07	3·53	1945 gold.
D/258	23349·32	7·21	3·09	1938 gold. Amalgam rings
D/117	23738·42	9·04	3·81	1938 gold. Annealed
C/251	25454·40	12·25	4·81	1938 gold.
C/276	25742·35	11·26	4·37	1938 gold.
D/269	29847·74	10·98	3·68	1938 gold.
D/261	30282·10	12·72	4·20	1945 gold.

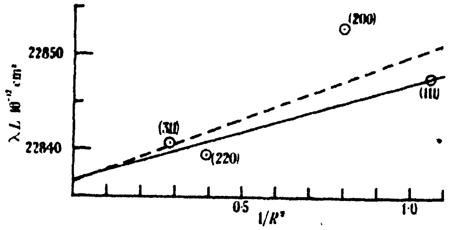


Figure 1.

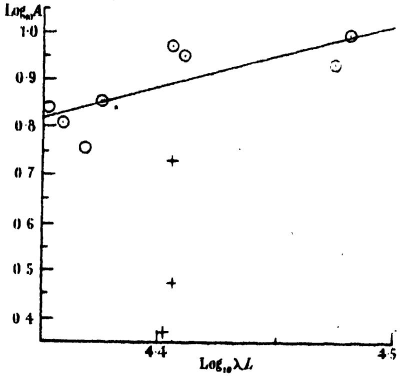


Figure 2.

shown in figure 2, which is a graph of the logarithm of the anomaly against the logarithm of the corresponding $(\lambda L)_0$. The straight line has been fitted by the method of least squares and has the equation

$$\log_{10} A = -5{\cdot}174 + (1{\cdot}40 \pm 0{\cdot}42) \log_{10} (\lambda L)_0. \qquad \ldots\ldots(12)$$

Within the limits of experimental error, the anomaly is *proportional* to $(\lambda L)_0$. Values for the ratio of the two are listed in the fourth column of table 5; their mean value is

$$\text{Anomaly}/(\lambda L)_0 = (3{\cdot}92 \pm 0{\cdot}20) \times 10^{-4}. \qquad \ldots\ldots(13)$$

The gold specimens were prepared from two samples of gold leaf purchased in 1938 and 1945 respectively. Comparison of table 5 with figure 2 shows that there is no perceptible difference between the results for the two samples. An analysis of the latter sample showed it to contain 0·09% copper and 0·02% silver. As it is not possible to obtain gold leaf of greater purity than this, the effect of impurity was examined by obtaining diffraction patterns from a less pure gold leaf containing 2·55% silver and 1·35% copper. The results for plates from this are given in table 6 and are plotted in figure 2 as crosses. It is clear that the addition of impurity, far from being the cause of the anomaly, actually tends to reduce it.

Table 6. Results for impure gold leaf

Plate No.	$(\lambda L)_0$ (10^{-12} cm^2)	Anomaly (10^{-12} cm^2)
D/293	25534·2	3·7
D/294	25290·8	3·0
D/305	25460·4	7·0

We have attempted to investigate this point further by obtaining diffraction patterns from gold films prepared by electro-deposition by the method of Finch and Sun (1936); such films might be expected to be more free from impurities than the best commercial gold leaf. There was an indication that the (200) anomaly was somewhat larger for such specimens, but it was impossible to be certain as the crucial (200) ring was always weak and the determination of the correction for background density consequently liable to large error.

The majority of the results of table 5 and figure 2 were obtained from gold leaf which had not been annealed. Plate D/117, however, was obtained from a specimen which had been annealed at 340° c. for 22 hours after thinning in potassium cyanide solution. Since the result for this plate is not sensibly different from that of the others, we may conclude that the anomaly is not due to any strain in the crystal which can be removed by annealing at this temperature.

The majority of the plates showed no trace of any amalgam rings (figure 5). However, two plates showed marked amalgam rings (figure 6). Since the values of the anomaly for these plates are not perceptibly different from its values for the remaining plates, we may conclude that traces of amalgamation can hardly be responsible.

It may be concluded that the (200) diffraction anomaly observed with gold-leaf specimens is a feature of the pure gold lattice, and that it does not arise from any strains which can be removed by annealing. It would have been of interest to examine specimens of gold evaporated on to cellulose; unfortunately, the diffraction rings obtained from such specimens are so broad that it is impossible to make measurements of sufficiently high precision.

As an additional test, measurements have been made on patterns from a specimen the plane of which was not perpendicular to the electron beam. Table 7 shows that such a tilted specimen yields a value for the ratio of the anomaly to $(\lambda L)_0$ which is not sensibly different from the average value (3·92) for the other plates.

Table 7. Effect of tilting specimen

Plate No.	Angle of tilt (°)	$(\lambda L)_0$ (10^{-12} cm^2)	Anomaly (10^{-12} cm^2)	$\frac{\text{Anomaly}}{(\lambda L)_0}$
D/280	65	23126·0	7·75	$3{\cdot}35\times10^{-4}$

In this section, we have regarded our results as indicating that the value of (λL) calculated from the (200) diffraction ring is higher than would be expected. Consideration of figure 1 shows that this is not the only possible explanation. The small difference in the abscissae of the points representing the (220) and (311) rings makes it equally possible to regard these two and the (200) point as lying on a straight line, in which case the value of λL for the (111) ring is abnormally low. According to the theory presented in the next section, there is no essential difference between these two interpretations.

§ 5. INTERPRETATION OF ANOMALY

The anomalies discussed in the preceding section might be attributed to one or more of the following causes: (*a*) refraction of the electrons at the surface of the gold crystallites, (*b*) variation of the angle of diffraction from the Bragg value in accordance with the dynamical theory, (*c*) deformation of the crystals from exact cubic form.

(*a*) The effect of refraction can be shown to be two-fold: the rings are broadened, or even split into doublets (Sturkey and Frevel, 1945), and the peak of the broadened ring is displaced. Such a displacement can be shown to result in an anomaly in the value of λL proportional to the *cube* of the electron wave-length. Now the experimental results as represented by equation (12) and figure 2 are consistent with a first-power law but definitely rule out the possibility of a cube law. The origin of the anomaly cannot therefore be sought in refraction effects.

(*b*) The fact that the intensities of the (111) and (200) rings of gold are considerably less than is required by the kinematic theory (Tol and Ornstein, 1940) suggests that dynamical effects are very marked. Nevertheless, the following considerations indicate that such effects are probably not the main cause of the observed anomalies. In the first place, Thomson and Blackman (1939) have shown that for transmission through a parallel-sided slab of crystal ("Laue case") the dynamical theory predicts a change in the angle of diffraction by an amount $\zeta\lambda \tan\psi/2\pi$, where ζ is the *resonance error* of Bethe's theory, λ the wave-length of the electrons and ψ the angle between the reflecting plane and the surface of the crystal. According to the dynamical theory, ζ ranges between approximately

$$\pm \frac{v\lambda}{2\pi},$$

where v is the Fourier coefficient corresponding to the diffraction, so that the angle of diffraction can deviate from the normal by amounts up to

$$\pm \frac{v\lambda^2 \tan\psi}{4\pi^2}.$$

The elementary dynamical theory therefore predicts no change in the mean radius of the diffraction ring but only a *symmetrical broadening*, which should be of the order of 200 μ for a camera length of 50 cm. Nevertheless, since the observed anomalies amount to only some 4% of the ring width, it is not inconceivable that a refinement of the present dynamical theory might account for them as a higher-order effect. Since the elementary dynamical theory leads to symmetrical deviations in the angle of diffraction proportional to λ^2, it would be expected that any such higher-order effect would be proportional to at least the second power of λ, while the experimental results (equation (12)) show that even a second-power dependence on λ is less likely than a first power, and any higher power than the second is definitely ruled out. We conclude that the possibility of the anomalies being due to dynamical effects cannot be entirely ignored, but that it is not very likely, and in any case cannot be further discussed, in the present state of the theory.

(*c*) If the anomalies are due to a departure of the crystals from cubic symmetry, they would be proportional to the *first* power of the electron wave-length, in agreement with equation (12). However, the assumption of a non-cubic lattice is not, of itself, sufficient to explain the results. If, for example, it be supposed that the crystal lattice is slightly deformed from cubic to tetragonal form without change of volume of the unit cell, then one of the (200) spacings is increased (decreased) while the other two are decreased (increased). The diffraction pattern from a random arrangement of such crystals would give a broadened

(200) ring but with its peak undisplaced. Similar reasoning can be applied to other planes of the crystal and to other assumed deviations from cubic form. The only way in which a pseudo-cubic crystal lattice could give the observed results is if there is some orientation of the crystallites so that—to take the example just given—the (200) planes of increased spacing are never approximately parallel to the electron beam. We must, however, reject this explanation for two reasons: (i) the degree of orientation as judged by the intensity of the rings and the amount of "arcing" when the specimen is tilted is small and varies considerably in magnitude from specimen to specimen, whereas the magnitude of the anomaly is consistent from one specimen to another; (ii) inclining the specimen to the electron beam makes no appreciable change in the value of the anomaly (see table 7). It therefore appears impossible to explain the results by assuming an ordinary unstressed crystal lattice with flat crystal planes; we are driven to postulate a *stressed* lattice in which the planes have been warped. Such a warping will generally occur when a gold crystal is stressed owing to the fact that it is elastically anisotropic.

While the evidence thus points to stresses in the gold crystals as the cause of the observed anomalies, it does not suffice to determine unambiguously the nature and origin of these stresses; it is found that stress systems of a rather general type are capable of explaining quantitatively the observed results.

Consider the effect of a *simple tension* p in the specimen in a direction making an angle ϕ with the electron beam. Referred to the crystal axes of a certain crystallite, let $\gamma_1, \gamma_2, \gamma_3$ be the direction-cosines of the electron beam and C_1, C_2, C_3 those of the normal to a set of reflecting planes. Then the fractional extension of the crystal along C_1, C_2, C_3 is:

$$\delta = p\left[\frac{-c_{12}}{(c_{11}+2c_{12})(c_{11}-c_{12})} + \frac{C_1^2\gamma_1^2 + C_2^2\gamma_2^2 + C_3^2\gamma_3^2}{c_{11}-c_{12}} + \frac{C_1C_2\gamma_1\gamma_2 + C_2C_3\gamma_2\gamma_3 + C_3C_1\gamma_3\gamma_1}{c_{44}}\right], \qquad \ldots\ldots(14)$$

where c_{11}, c_{12}, c_{44} are the usual elastic constants.*

The fractional extension along the [111] direction of the crystals producing the (111) diffraction ring may be found as follows. We have

$$C_1 = C_2 = C_3 = 1/\sqrt{3}.$$

Hence

$$C_1^2\gamma_1^2 + C_2^2\gamma_2^2 + C_3^2\gamma_3^2 = \tfrac{1}{3}. \qquad \ldots\ldots(15)$$

Now let the normal to the (111) planes of a diffracting crystal (this normal must necessarily be very nearly perpendicular to the electron beam) make an angle θ with the projection of the tension p on to a plane perpendicular to the electron beam; the corresponding portion of the (111) diffraction ring is in azimuth θ with respect to the projection of p on the photographic plate. Then the angle between p and the [111] direction is $\cos^{-1}(\sin\phi\cos\theta)$. Hence

$$C_1\gamma_1 + C_2\gamma_2 + C_3\gamma_3 = \sin\phi\cos\theta.$$

whence from (15)

$$C_1C_2\gamma_1\gamma_2 + C_2C_3\gamma_2\gamma_3 + C_3C_1\gamma_3\gamma_1 = \tfrac{1}{4}\sin^2\phi - \tfrac{1}{6} + \tfrac{1}{4}\sin^2\phi\cos 2\theta. \qquad \ldots\ldots(16)$$

---

* *Handbuch der Physik*, 1928, **6**, 418 (Berlin: Springer).

Substituting from (15) and (16) in (14),

$$\delta = p\left[A + \tfrac{1}{3}B + \frac{1}{4c_{44}}C\right], \qquad \ldots\ldots(17.1)$$

where

$$A = \frac{-c_{12}}{(c_{11}+2c_{12})(c_{11}-c_{12})} + \frac{\sin^2\phi}{2(c_{11}-c_{12})},$$

$$B = (1-\tfrac{3}{2}\sin^2\phi)\left(\frac{1}{c_{11}-c_{12}} - \frac{1}{2c_{44}}\right),$$

$$C = \sin^2\phi\cos 2\theta.$$

By an extension of this method, it can be shown that the corresponding expressions for the other rings are all of the form

$$\delta = p[A + \alpha B + \beta C], \qquad \ldots\ldots(17.2)$$

where A, B and C are the same for all rings. The coefficients α and β for the first four rings are given in table 8.

Table 8

Ring indices	$1/(h^2+k^2+l^2)$	α	β
111	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{4c_{44}}$
200	$\frac{1}{4}$	0	$\frac{1}{2(c_{11}-c_{12})}$
220	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8(c_{11}-c_{12})}+\frac{3}{16c_{44}}$
311	$\frac{1}{11}$	$\frac{19}{121}$	$\frac{32}{121(c_{11}-c_{12})}+\frac{57}{484c_{44}}$

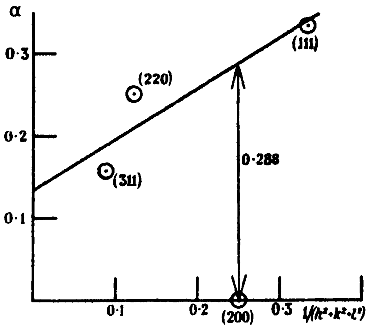


Figure 3.

Now the first term of (17) represents a uniform extension of the crystals in all directions and corresponds merely to a change of scale of the diffraction pattern; it is therefore irrelevant for the present results. The last term gives rise to an ellipticity of the rings, and will be considered later. The second term vanishes in the case of elastically isotropic crystals for which the Cauchy relation $c_{11}-c_{12}=2c_{44}$ is valid, while for anisotropic crystals it gives rise to changes in the relative radii of the different diffraction rings.

If a certain interplanar spacing is *increased* by a small fraction δ, the radius of the corresponding diffraction ring is *decreased* by δ and the value of (λL) computed from it by means of equation (6) using the normal interplanar spacing is likewise decreased by a fraction δ. Hence the ordinates of a $(\lambda L, 1/R^2)$ graph such as figure 1 are displaced by amounts proportional to $-\alpha$. The abscissae of figure 1 are the quantities $1/R^2$, which are very approximately proportional to d^2, i.e. to the numbers in the second column of table 8. The anomaly in the (200) reflection may therefore be evaluated by plotting α against $1/(h^2+k^2+l^2)$ (figure 3), fitting the best straight line to the points corresponding to the (111), (200) and (311) reflections and observing the distance of the (200) point from this line. It is found that this displacement is 0·288. Comparing this with the observed value of (anomaly)/λL (equation 13), we get

$$0{\cdot}288pB = 3{\cdot}92 \times 10^{-4}$$

whence

$$p(1-\tfrac{3}{2}\sin^2\phi)\left(\frac{1}{c_{11}-c_{12}}-\frac{1}{2c_{44}}\right) = 1{\cdot}36\times 10^{-3}.$$

Inserting the known values of the elastic constants (Goens and Weerts, 1936),

$$c_{11} = 18{\cdot}6\times 10^{11} \text{ dynes/cm}^2$$
$$c_{12} = 15{\cdot}7,$$
$$c_{44} = 4{\cdot}2,$$

we obtain

$$p(1-\tfrac{3}{2}\sin^2\phi) = 6{\cdot}01\times 10^{8} \text{ dynes/cm}^2$$

The anomaly can therefore be explained if we suppose that there is a tension of 6·01 × 10^8 dynes/cm^2 parallel to the electron beam ($\phi = 0$); such a tension causes a displacement of the (200) point from the line fitting the other three points, while the displacements of the latter are negligible.

It should be noted that we cannot compare the *slope* of the line of figure 3 with the slope of the experimental curve of figure 1, for it is known (Rymer and Butler, 1945 b) that the slope of the latter is in part due to a charging up of the photographic plate by the undiffracted beam. The magnitude of this charging-up varies from one photograph to another and cannot therefore be easily allowed for. Its effect is to add to the ordinates of the points quantities proportional to $1/R^2$; it therefore cannot change the magnitude of the anomaly.

The stress system postulated above is not a unique solution to the problem, for since a uniform hydrostatic compression reduces *all* interplanar spacings by the same fraction,* a stress system consisting of a simple tension together with an arbitrary hydrostatic pressure will fit the experimental results equally well. In particular, a stress system consisting of a tension p and a hydrostatic compression of the same magnitude is a possible solution. This reduces to a two-dimensional compression in a plane perpendicular to the electron beam. Such a stress system could arise from surface-tension forces if the specimen is in the form of laminae set normal to the beam (i.e. in the plane of the specimen); the interior of a lamina of thickness t of material with a surface tension S will experience a compression in its plane of magnitude $2S/t$. There is no information as to the surface tension of solid gold, and the published values for the molten metal range from 500 to 1000 dynes/cm. If we take the former value, we find $t = 1{\cdot}7\times 10^{-6}$ cm. as the thickness of a lamina. This is of the order of magnitude of the thickness of transmission specimens.

Equation (17) predicts that when the laminae are not perpendicular to the electron beam ($\phi \neq 0$), two effects should be observed: (i) the magnitude of the anomaly, which is proportional to B, should be reduced by a factor $1-\tfrac{3}{2}\sin^2\phi$; (ii) the term βC no longer vanishes, indicating that the rings become elliptical. However, the results of table 7 show that inclining the specimen to the beam does

---

* *Handbuch der Physik*, 1928, **6**, 418 (Berlin: Springer).

not change the value of the anomaly. Investigation of the ellipticity of the rings is hampered by the unavoidable presence of stray magnetic fields, to which reference has been made in § 3, but the results (which need not be given in detail) show that the ellipticity attributable to strain is not statistically significant, and in any event is smaller by a factor of 10 than is predicted by (ii). These results can be brought into harmony with the stress theory if it be supposed that in the region of the specimen irradiated by the beam there is a large number of domains with a different direction of the tension in each. The ellipticity of the rings would be averaged out, while the (200) anomaly would be that corresponding to the average value of $\sin^2\phi$. Such an effect could be produced by surface-tension forces if the specimen consisted of a number of laminae in random orientation and having a thickness of the order of 4×10^{-7} cm. (assuming a surface tension of 500 dynes/cm.). A specimen in the form of filaments of radius of this order and in random orientation would equally give rise to the observed effects.

It is apparent from this discussion that the results are consistent with a wide variety of stress systems, and it might therefore be expected that the stresses in the neighbourhood of lattice imperfections such as dislocations or twinning planes would give rise to the observed effects. There are, however, two difficulties in attributing the results to this cause. First, in the neighbourhood of a lattice imperfection there are two regions of equal and opposite stress, and the diffraction rings from these would be displaced by equal and opposite amounts: the resultant ring would be slightly broadened but would not be displaced. Secondly, it would be expected that the neighbourhood of a dislocation would be characterized by a *strain* which would not be sensitive to small traces of impurity, for the essential feature of a dislocation is that a group of atoms is displaced through a distance determined by the lattice constant. The *stress* associated with a dislocation will of course be sensitive to traces of impurity. Now it is found experimentally (compare tables 5 and 6) that the addition of 2·6% of silver greatly reduces the (200) anomaly. This implies a *stress* which is independent of traces of impurity and a *strain* which is diminished when the lattice is hardened by the presence of foreign atoms. This is consistent with a surface tension rather than a dislocation origin of the stress system. Another fact pointing in the same direction is that the magnitude of the (200) anomaly is unchanged by annealing, though this is not conclusive as we were unable to use temperatures above 340° c. without risk of damage to our specimens.

A surface-tension origin of the stress system means that the magnitude of the (200) anomaly depends on the thickness of the diffracting particles, whereas the results of table 5 show that it is very consistent from one specimen to another. It is to be expected that particles of a wide range of thickness will be present. Particles thicker than a certain amount will contribute little to the pattern owing to excessive absorption of the beam. Considerations of mechanical strength will set a lower limit to the size of the particles, and also the smallest particles will have insufficient scattering power to produce a good pattern. The bulk of the diffraction pattern will therefore come from particles of a rather restricted range of thickness, and this probably accounts for the consistency of the observed anomaly from one specimen to another.

§6. ACKNOWLEDGMENTS

The authors gratefully acknowledge the encouragement they have received from Professor J. A. Crowther and the facilities placed at their disposal. The senior author is also indebted to the Research Board of the University of Reading for a grant for the purchase of the electron-diffraction camera.

REFERENCES

Becker, A. and Kipphan, E., 1931. *Ann. Phys., Lpz.*, **10**, 15.
Finch, G. I., Quarrell, A. G. and Wilman, H., 1935. *Trans. Faraday Soc.*, **31**, 1051.
Finch, G. I. and Sun, C. H., 1936. *Trans. Faraday Soc.*, **32**, 852.
Goens, E. and Weerts, J., 1936. *Phys. Z.*, **37**, 321.
Rymer, T. B. and Butler, C. C., 1944. *Phil. Mag.*, **35**, 202 ; 1945 a. *Ibid.*, **36**, 515 ; 1945 b. *Ibid.*, **36**, 821.
Sturkey, L. and Frevel, L. K., 1945. *Phys. Rev.*, **68**, 56.
Thomson, G. P. and Blackman, M., 1939. *Proc. Phys. Soc.*, **51**, 425.
Tol, T. and Ornstein, L. S., 1940. *Physica*, **7**, 685.

A COLORIMETER WITH SIX MATCHING STIMULI

By R. DONALDSON,
National Physical Laboratory, Teddington

MS. received 19 *November* 1946

ABSTRACT. The instrument is a modification of the ordinary trichromatic colorimeter. The three matching stimuli of the ordinary instrument, the red, green and blue, have been increased to six by the addition of an orange, yellow-green and blue-green. The spectral energy distribution of the colour being measured is first approximately matched by means of a mixture of all six colours before the final colour match is made by varying three of the colours only. This eliminates to a large extent the personal error of the observer, and allows a large field to be used with a resultant gain in sensitivity.

§1. INTRODUCTION

In the measurement of colour there is naturally a tendency to pass from visual to photoelectric methods. This change-over, however, is not taking place as smoothly as might be expected. The difficulty is that there has not yet appeared a simple photoelectric design which will permit the construction of a cheap reliable instrument. There have been two main lines of development—the spectrophotometer, and the photoelectric colorimeter employing a spectrum template. The former would seem, for the present, to have reached a culmination in the Hardy automatic instrument, and the latter, although it has not received serious attention for such a long time as the spectrophotometer, has already produced two versions of promise (Knipe and Reid, 1943; Winch, 1946).

When considering the obvious advantages of photoelectric methods, the considerable increase of complexity in the apparatus must not be overlooked. A photoelectric instrument, which measures colour accurately and quickly, is

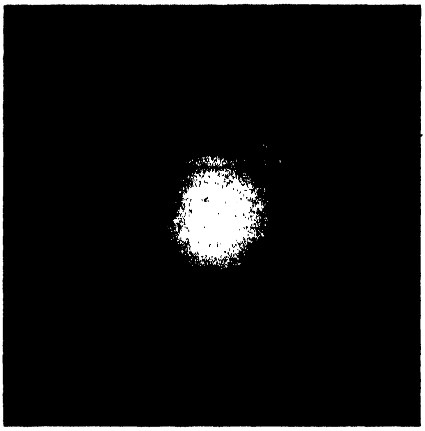


Figure 4. Plate D/137.
Sodium chloride specimen.

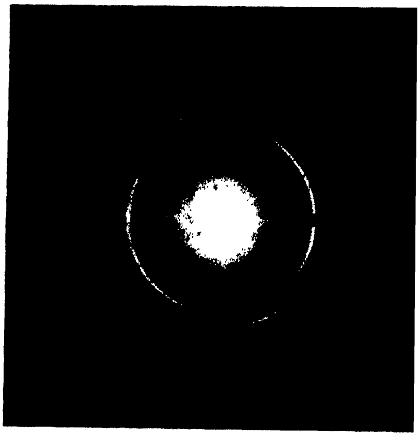


Figure 5. Plate D/268.
Gold-leaf specimen.

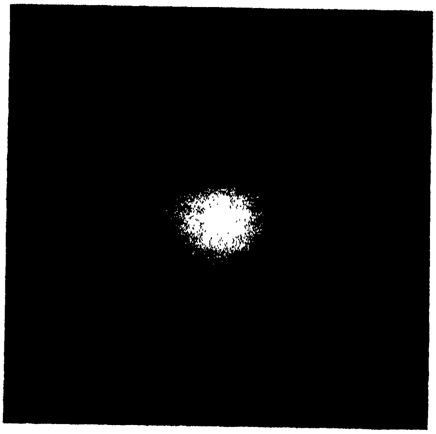


Figure 6. Plate D/258.
Gold-leaf specimen showing amalgam rings.

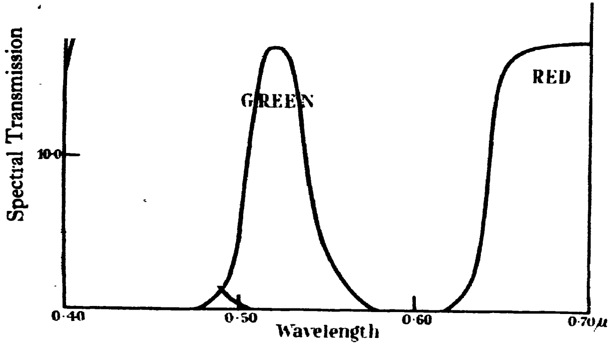


Figure 1. Mixture colours of the ordinary trichromatic colorimeter

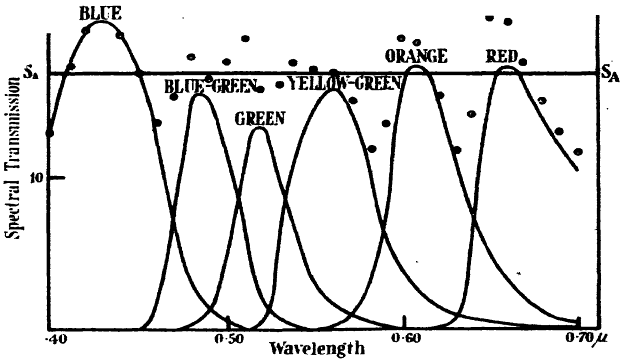


Figure 2. Energy match with standard illuminant A.

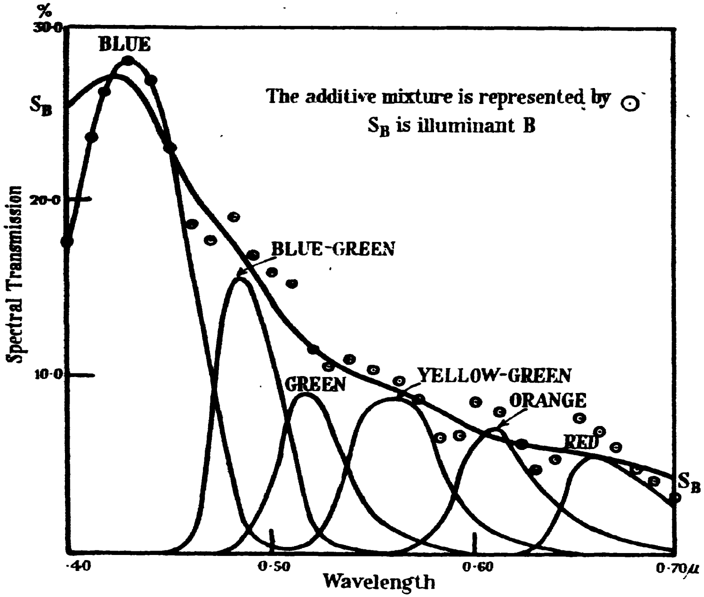


Figure 3. Energy match with standard illuminant B

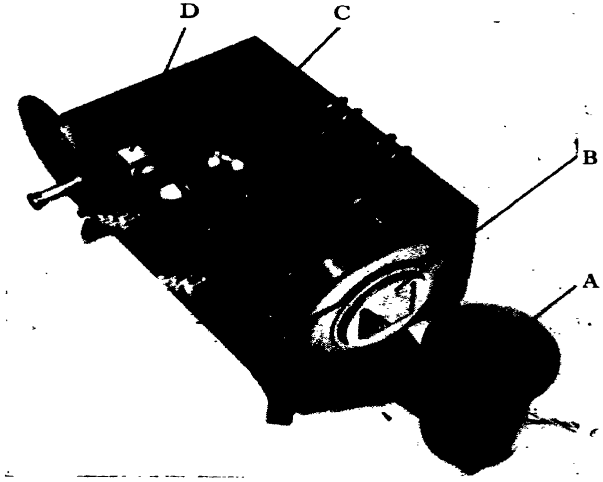


Figure 4. Colorimeter with six matching stimuli.

elaborate and expensive, and there seems little prospect of improvement in this direction. The visual instrument is, in general, more robust and requires less maintenance. It is consequently suited to the unskilled or semi-skilled observer who, once the routine has been learned, can carry on with little skilled supervision.

There is another natural advantage of visual observation which should be mentioned, for it is missed very much when going over to photo-electric methods, that is, its sensitivity to low brightnesses. A visual instrument can be made to measure practically anything that can be seen. In addition to the measurement of filters and reflecting colours with the standard illuminants A, B and C, it can deal conveniently with any kind of illuminant and coloured specimens under that illuminant. The measurements can be made at ordinary levels of illumination, and it is not necessary to arrange for artificially high values of illumination to get the desired accuracy. When using direct observation of a tungsten filament it is possible to measure even the densest welders' protective goggles, which can reach an optical density of 6. It is for such reasons that visual methods cannot yet be regarded as superseded, but still have an important part to play in colour measurement.

In the following design of visual colorimeter two of the main defects of the ordinary trichromatic colorimeter have been removed: firstly, the large personal error of the observer, and secondly, the lack of sensitivity due to the rather small field. These improvements have introduced a little more complication in the instrument itself, and also a longer calculation in transforming the results. The observational work is reduced, however, so that, as far as the time for a complete measurement is concerned, there is an even balance between the two types of instrument.

§2. PERSONAL ERROR OF OBSERVER

This instrument can be regarded as an extension of the ordinary sphere colorimeter (Donaldson, 1935) with three mixing colours. As is well known, everyone can get a perfect colour match with three colours, but the settings vary with the observer. From the point of view of measurement this is a serious defect. Two different observers can, in measuring certain colours, get widely different results, although all their observations are closely grouped about their respective mean values. The cause of this is the combination of the observer's personal colour-vision characteristics and the differences of spectral energy distribution between the colours being matched. The colour being measured has in general a continuous distribution, whereas the instrument colour is a mixture of red, green and blue spectral bands only. In figure 1 are shown the spectral transmissions of the trichromatic instrument filters, which, when illuminated by illuminant A, form the instrument stimuli. It can be seen from figure 1 that there are big gaps in the energy distribution of the instrument colour.

In the present instrument three more mixing colours have been added, so that these gaps are filled in and the instrument colour is made to resemble more closely the colour being measured. A blue-green is inserted between the blue and the green, and a yellow-green and orange between the red and the green, this being the bigger gap. The filters were chosen to fit into each other as smoothly as possible so that the fall in transmission on one side of a filter is counterbalanced by the rise in an adjacent filter. In figure 2 are shown the spectral transmissions of the six filters and also the kind of fit with illuminant A when they are additively mixed together. Figure 3 shows the fit with illuminant B. For other smooth distributions, a similar order of fit is obtained.

Details of the construction of the filters

Red: Chance OR1, 1·8 mm., and Calorex, 3·3 mm.
Orange: Chance OR2, 2·5 mm., Corning 978, 2·9 mm., and cadmium yellow, 0·9 mm.
Yellow-green: Chance OGr1, 2·8 mm., and cadmium yellows, 1·7 mm. and 1·4 mm.
Green: Chance OY4, 2·1 mm., and Zeiss BG7, 4·2 mm.
Blue-green: Wratten gelatine filter No. 75.
Blue: Chance OB1, 2·5 mm.

The cadmium yellows are unlisted yellow glasses in common use for signal glasses and fog lamps etc. They can be duplicated from Chance's later catalogues. It was found impossible to construct a suitable glass filter for the blue-green stimulus.

With three more mixing colours, there are six controls on the instrument and consequently there is no longer a unique setting for each colour match. The question therefore arises as to how the controls should be set so that there is an approximate energy match to the unknown colour that is being measured. The procedure for this is as follows. To set the red stimulus, a red filter of the same nature as the instrument filter is held at the eye and the red control is varied until there is a brightness match in the field. The process is repeated with an orange filter and the orange stimulus, and so on for each of the filters in turn. Owing to the slight overlap of the filters we require to repeat the process a second time, but very soon a state is reached where the instrument colour is in agreement with the colour being measured when viewed through each of the six filters in turn.

When the controls have been set in this way, there is in general an approximate but not an exact colour match in the field. To get an exact colour match three controls only—red, green and blue or orange, green and blue—are adjusted in the usual way. The amount of adjustment is so small that it does not disturb the energy match appreciably. At the matching point, therefore, the observer is only asked to discriminate between two colours of nearly the same energy distribution. Under these conditions there are no wide differences in the settings with different observers, and consequently the personal error is considerably reduced.

§ 3. CONSTRUCTIONAL

The mechanical construction follows very closely that of the earlier sphere colorimeter (Donaldson, 1935). The linear scales of the stimuli are produced by apertures with sliding shutters and the colour mixing is carried out by an integrating sphere as in that instrument. The accompanying photograph, figure 4, shows the arrangement in the interior of the six-stimuli colorimeter, lamp A, condensing lens and apertures B, integrating sphere C and the photometric cube D. There are six rectangular apertures in front of the condensing lens. They are arranged in two columns, three on each side, with the sliding shutters opening outwards from the centre. The scales are engraved on the inside of the shutters and read by means of lenses and mirrors via the interior of the instrument. The filters are mounted on the outside. This is preferable to having them behind the shutters because in the outside position they are uniformly heated by the lamp.

It is important to have a smooth motion on the shutters without backlash. Transmission cables in tubes were used for three of the controls but, as they are not quite successful, pulleys and strings were fitted to the other three. The latter have proved to be quite satisfactory and provide a movement that feels pleasantly smooth and direct.

§ 4. FIELD SIZE

The removal of the energy differences between the colours being matched allows complete freedom in the choice of field size. In the trichromatic colorimeter the 2° field is standard. This size was adopted to ensure freedom from Purkinje effect over a large range of brightness and also to be in agreement with the standard observing conditions under which the response curves describing the colour and luminosity functions of the normal observer have been obtained. The practical need for this restriction only arises when there are appreciable differences in the energy distributions of the matched colours.

When the energy differences are removed or partially removed, as in this instrument, there is no need to restrict the size of the field. Large fields are more sensitive to colour differences than the small 2° field. A field of 15° angular size has therefore been chosen and the Lummer-Brodhun contrast patches have also been added. The Lummer-Brodhun field allows the eye to attain practically its limit of colour sensitivity. Small colour-differences which can be just seen under ordinary viewing always seem to be enhanced in the Lummer-Brodhun field. In colour-temperature work an accuracy of $\pm \frac{1}{2}\%$ in volts can easily be obtained with it. This corresponds to a maximum change of about 0·0005 in the trichromatic coefficients. The equivalent of this high discrimination is probably maintained over the whole of the colour field. As a result it is almost impossible to get a colour match on the instrument that is completely satisfying when looked at critically. Sufficient accuracy for all practical purposes can be obtained, however, by three or four quite casual matches. Matching casually saves a great deal of eye-strain. It is found with the majority of ordinary specimens that the variations due to non-uniformity generally tend to be greater than the smallest differences discernible in the field, so that as far as colour sensitivity is concerned the Lummer-Brodhun field is adequate.

§ 5. TRANSFORMATION EQUATIONS

The results as given by the instrument are arbitrary readings, in terms of scale divisions of red, orange, etc. A set of equations is therefore required which will transform to the C.I.E. standard reference stimuli X, Y and Z. In deriving the equations, two aspects of each mixing colour have to be defined. There is the colour quality, or *chromaticity*, and the amount that is present in the mixture. The colour quality is found by the usual method of spectrophotometry and calculation. The quantities of red, orange, etc., which correspond to a scale division cannot be obtained so directly. In the ordinary trichromatic colorimeter, these quantities are defined by means of a colour match made with white, usually standard illuminant B. With six colours there are too many to be related to each other by colour matching, but they can be related by a series of brightness matches. The auxiliary filters used to analyse the spectral distribution of the colour being measured also serve as standards for the brightness matches. In this measurement they are not placed at the eye but in the usual position for the measurement of transparent specimens and illuminated by illuminant A. They are of the same colour and energy distribution as the respective instrument stimuli so that the results are independent of the observer's colour vision. The transmission of each auxiliary filter is known, so the quantities of the matching stimuli can be related to each other and the transformation equations derived. The transformation equations are six in number, and a typical example is as follows :

$$\begin{aligned}
R &= 1{\cdot}847X+0{\cdot}696Y+0{\cdot}000Z\\
O &= 11{\cdot}387X+5{\cdot}933Y+0{\cdot}005Z\\
YG &= 5{\cdot}943X+8{\cdot}031Y+0{\cdot}073Z\\
G &= 0{\cdot}730X+3{\cdot}360Y+0{\cdot}435Z\\
BG &= 0{\cdot}200X+0{\cdot}849Y+1{\cdot}652Z\\
B &= 1{\cdot}167X+0{\cdot}197Y+6{\cdot}231Z
\end{aligned}$$

The quantities R, O, etc., refer to one unit of each of the matching stimuli, one scale division of red, orange, etc. The right-hand side of each equation is proportional to the trichromatic coefficients defining the chromaticity of the instrument stimulus. The proportions are such that the ratios of the coefficients of Y are as the relative luminosities of one division of red, orange, etc. The method of using the equations is the same as that for the ordinary trichromatic transformation.

There is one important difference, however, between sets of equations derived in this way and by the method used in the trichromatic colorimeter, i.e. the white is no longer automatically given the correct value. The measurement of white is the same as for any other colour, and small experimental errors may appear in it. When measuring colours close to white, we can take advantage of the readily available standard, magnesium oxide, and use the instrument as a differential colorimeter to measure the difference between the standard and the near-white. This difference, if small, will be free from any systematic error due to the instrument.

§ 6. ACCURACY

The ideal measuring instrument should give results conforming to the average observer when used by ordinary observers having the usual variations in colour vision. These variations should not be capable of seriously upsetting the instrumental results. As the spectral energy matches in this instrument are never quite exact, the residual differences may cause some slight variation with observer. There are also the experimental errors in the determination of the constants of the instrument, e.g. in the colours of the filters, in the brightness matches required for the derivation of the transformation equations and in the setting of the templates controlling the linearity of the matching stimuli. All these factors taken together seem to have a greater influence on the experimental error than the chance variations in matching, which are very low on account of the high sensitivity of the Lummer-Brodhun field. This is shown by the fact that the repetition by a single observer is often better than the agreement with the normal observer.

It is also noticed sometimes that observers show a bias in a given direction with certain colours. This would seem to indicate that colour-vision variations are no longer the most important factor reducing the accuracy. There are the small, residual errors, due to the system of shutters and templates, and (what is probably more important) those due to the use of glass filters. Filters are never strictly uniform. They show variation in colour over their surface, and it needs very careful selection to keep this down to negligible proportions. Our experience of the instrument has shown that for the majority of colours the personal error of the observer has been reduced to the order of those inevitable errors arising from the mechanical construction of the instrument.

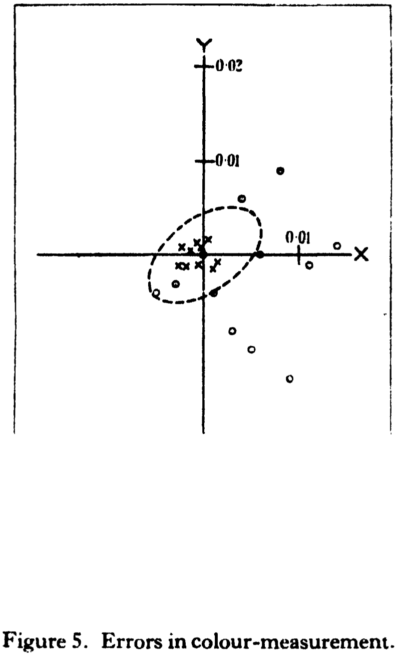


Figure 5. Errors in colour-measurement.
Yellow
$0{\cdot}6000X + 0{\cdot}3993Y + 0{\cdot}0007Z$
X 6-stimuli instrument.
⊙ 3-stimuli instrument.
- - - - MacAdam ellipse.

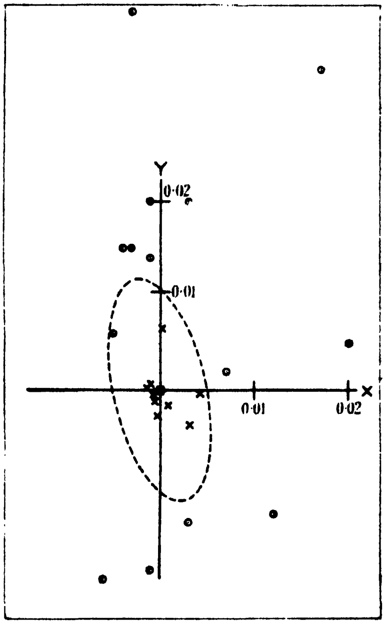


Figure 6. Errors in colour-measurement.
Blue-green
$0{\cdot}1918X + 0{\cdot}3976Y + 0{\cdot}4106Z$
X 6-stimuli instrument.
⊙ 3-stimuli instrument.
- - - - MacAdam ellipse.

In figures 5 and 6 are shown comparisons with this instrument and the ordinary trichromatic colorimeter of measurements made with two coloured filters, a yellow and a blue-green. These results refer to three observers and have been obtained at various times in the course of testing trichromatic colorimeters and the measurement of signal colours on the six-stimuli instrument. The origin of co-ordinates is the calculated colour of the filter. To give some indication of the colour sensitivity in these regions of the colour chart the MacAdam (1942) ellipses on a scale of three times the standard deviation have been sketched in. The ellipses represent a just noticeable colour difference. The improvement of the six-colour instrument over the three is very marked for these colours.

This instrument has been in service for some years now and it has been found that in general its accuracy with other colours is of the same order as shown on the diagrams.

§ 7. ACKNOWLEDGMENTS

The development of this instrument has been carried out as part of the research programme of the National Physical Laboratory, and this paper is published by permission of the Director of the Laboratory.

REFERENCES

Donaldson, R., 1935. *Proc. Phys. Soc.*, **47**, 1068.
Knipe, G. F. G. and Reid, J. B., 1943. *Proc. Phys. Soc.*, **55**, 81.
MacAdam, D. L., 1942. *J. Opt. Soc. Amer.*, **32**, 247.
Winch, G. T., 1946. *Trans. Illum. Engng. Soc., Lond.*, **11**, 107.

THE RECOGNITION OF COLOURED LIGHT SIGNALS WHICH ARE NEAR THE LIMIT OF VISIBILITY

By N. E. G. HILL,
Royal Aircraft Establishment, Farnborough, Hants

MS. received 24 *September* 1946

ABSTRACT. Statistical tests on the recognition of colour were made during 1938–39 to find the range of colours which would be best for aviation signals. Seventy-three colours were seen as point sources and viewed by binocular foveal vision, with dark-adapted eyes against a dark background, by nine observers of normal colour vision. The results were plotted as recognition contours, for eye illuminations of 1 mile-candle and 2 mile-candles respectively, on the x, y colour diagram for the colour categories red, yellow + orange, green + blue, and white. The results indicate that yellow + orange is the least satisfactory colour group for signals of low illumination. A modification is suggested to the specification for " aviation white " to avoid the risk of confusion with yellow + orange.

§ 1. INTRODUCTION

A characteristic feature of a coloured light signal is that its colour becomes less pronounced as the illumination of the signal at the observer's eye is reduced, and may disappear entirely before the limit of visibility is reached, so that at low values of illumination the chance of confusion between colours is increased. This effect is more marked with the paler or less saturated colours. In choosing colours for long-range light signals it is therefore necessary to select those colours which are the most recognizable when seen as point sources of low illumination. The choice of coloured signals has always been based on accumulated experience with particular colours, but it was thought, during the years before the war, that systematic data should be obtained on the recognition of coloured point sources in order that the full range of possible colours might be known. Data of this kind were obtained at the Royal Aircraft

Establishment during 1938 and early 1939 in an endeavour to find the best colours for aviation signals but, owing to the war- time restrictions, these data could not be published until now. Thus, although the data are not recent, they are new in the sense that they have not previously been published, and they are presented in the present paper in the belief that they may still serve as a contribution to the knowledge of colour recognition.

The tests here described were made with binocular foveal vision under conditions closely related to those under which aviation signals are usually observed but, in order to obtain consistent data which could be compared with those of other investigators, the conditions were idealized and closely controlled.

The background brightness was fixed at about that of starlit sky, the effects of atmospheric absorption were eliminated, and the tests were confined to the five groups of colour which are usually used for aviation signals, viz., red, yellow and orange, white, green, blue. Of these colours, blue is normally used only as a short-range signal, but the remainder are long-range colours. It was decided not to attempt to separate yellow recognition from orange because it seemed unlikely that this could be done satisfactorily at low values of eye illumination. Nor was any attempt made to obtain recognition figures for purple, which is known to be unsatisfactory at long range.

§ 2. THE PROBLEM

The recognition of a coloured light signal is a subjective reaction which in general cannot be predicted absolutely for a single observation, even for the average observer, but which can, however, be expressed as a probability for a single observation. For instance a particular signal, seen 100 times by an average observer under certain conditions specified, might be judged to be red 85 times, yellow 10 times, and white 5 times. The particular colour would therefore be recognized as red on 85% of the observations, and would have an 85% probability of being recognized as red on any single observation. Such a colour may be defined as having a red recognition of 0·85 under the specified conditions of observation. The basic problem in obtaining recognition data is thus a statistical one, and it was this consideration which governed the arrangements for the tests here described.

It was evident that a large number of observations of each signal would be required, and that each observation must be an independent one, unprejudiced by the judgment made on any previous observations of the same signal. The best solution to this problem appeared to be to present a succession of coloured signals, in random sequence, to an observer who was required to place each colour in one of a number of specific colour categories, and to repeat the process until sufficient observations were made. This method is similar to that used by McNicholas (1936) in a series of tests on signal glasses to determine the best set of six colours for use in a system of railway signals.

Some additional requirements had considerable influence on the apparatus and methods used. These were that the colours of the test signals should be spread as widely as possible over the colour diagram, that a number of normal observers should participate to an equal extent, that each signal should be observed a large number of times by each observer, and that random errors should be reduced by strict control of test conditions.

§ 3. THE COLOURED SIGNALS

The production of a large number of coloured point-signals, adequately spread over the possible range of colours, presented difficulties. A sufficient number of different single filters was not available, and the scheme of mixing coloured lights in various proportions was therefore considered.

The principle of the trichromatic colorimeter offered attractive possibilities both for the production of an adequate range of colours and for simplicity of control. The theoretical and experimental bases of the trichromatic theory have been clearly and adequately stated by Judd (1930) and Guild (1931) and summarized by Stiles, Bennett and Green (1937) and others, and are too well known to need further discussion here.

Three filters were chosen having, in conjunction with a standard illuminant, colours R, G, B, spaced widely over the colour diagram in figure 1. Light from these three filters was mixed in a diffusing sphere having a window covered by a pinhole. The quantity of light through each filter was controlled by means of a shutter and the colour and intensity of the illumination on the pinhole could thus be varied within wide limits. It will be clear from figure 1 that any colour C within the triangle RGB could be produced at the pinhole.

It was however found that this method of producing coloured point sources failed because, due to chromatic aberration in the eye, the component colours of the mixture were separated, and the apparent colour of the point was entirely changed. For instance, if the proportions of the primary colours were arranged to give what appeared a good white when the sphere window was viewed at short range, then, at long range, the pinhole appeared as a red point surrounded by a green-blue halo. It is clear that the light produced by mixing the three colours in the way described has a spectral composition which may be entirely different from that of light from a single filter of the same colour as the mixture. It was thought that satisfactory results would be obtained only if the spectral composition of the light used for the coloured point source was similar to that of the light from a single filter and light-source combination of the same colour.

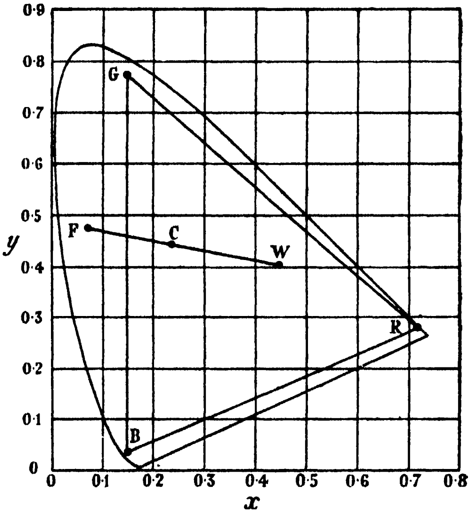


Figure 1. Colour diagram on 1931 I.C.I. standard reference system, showing method of obtaining colour mixtures.

The trichromatic method of producing coloured point sources was therefore abandoned and the alternative hue-and-saturation method was tried. A number of filters was obtained whose colours, in conjunction with a filament lamp, were represented by points as near the boundaries of the colour diagram as possible. Light from any one of these filters was mixed with white light in the diffusing sphere already referred to, and the mixture used to illuminate a pinhole. By varying the proportions of colour F and white W, a large range of colour mixtures C was made available (see figure 1).

The spectral distributions of such mixtures are similar to, though not identical with, the distribution of light from a single filter of the type encountered in practice. No difficulty due to eye aberration was experienced in viewing point sources formed in this way, and it was decided that, since the spectral transmission of the filters used was specified (1938), and since the spectral energy distribution of the point sources could thus be calculated if desired, recognition tests could usefully be made.

§4. DESCRIPTION OF APPARATUS

It was clear that, as a very large number of separate observations would be required, great care would have to be taken to ensure consistent reproduction of each signal and to avoid tiring the observers. The test apparatus was therefore designed with a view to ensuring the accurate presentation of each colour, and rapid change from colour to colour.

The apparatus is represented in diagrammatic form in figure 2. The coloured filters were mounted on a vertical disc (2), and light from the lantern (1) passed through a filter and the clear glass sheet (3) into the diffusing sphere (5). The light source consisted of a 200-v., 500-w. class A1 projector lamp backed by a plane silvered-glass mirror. The filter disc was arranged to rotate and was provided with a ratchet so that any desired filter could be quickly and accurately brought into position. The "white light" source (4) consisted of a 12-v., 60-w. motor-car type lamp, also backed by a plane silvered-glass mirror. The light from this source was reflected from the clear glass sheet into the diffusing sphere for mixing with the coloured light. Each lantern was arranged on slides along its light axis and was provided with an index and calibration scale, the two sets of slides being at right angles.

The details of the diffusing sphere are shown in figure 3. The light entered the sphere through the larger opening and was prevented from passing right through by one central flat screen. The inside of the sphere and the screen were silver plated, polished, and then coated with a uniform layer of magnesium oxide. The coloured and white lights were completely mixed inside the sphere, and the composite light emerged from the smaller opening of the sphere, outside which was placed a pinhole of 0·0496 inch diameter.

The pinhole, which acted as a luminous coloured point source, was viewed by the observer seated at 24·5 feet distance. The position of the observer's eyes was fixed by a binocular eyepiece which could be adjusted to suit the distance between the eyes, and which did not restrict the pupil.

The wide range of transmissions of the filters used for the tests necessitated the provision of a variable sector disc (7), figure 2. The point source was made visible to the observer by means of a rotary shutter on the flasher unit (8), which was controlled electrically and arranged to give a single flash of definite duration.

It was found desirable to provide a means of focusing the observer's eyes on the correct spot before the flash of the point source occurred. If this was not done, a considerable part of the flash was wasted while the eye searched for the point and then focused on it. Accordingly light from a 12-v., 4-w. lamp (10), operating at 8 volts, was allowed to fall on the black surface of the flash shutter for about 1½ seconds immediately prior to the flash. This preliminary light appeared to the observer as a dim area of illumination covering the aperture in the screen (9); the angular diameter of the area was about 15 minutes of arc, and its brightness of the order of 0·0005 candles per square foot; this was found to be very suitable for the purpose and formed an essential part of the test apparatus. The preliminary light, because of its low brightness and of its dissimilarity in character from the main signal, did not prejudice the observer in his opinion of the colour of the point source. As an additional help to the observer, a warning gong was sounded once about ½ second before the preliminary light appeared. The sequence was initiated by a push button and operated through a system of relays controlled by cam contacts on the flasher motor.

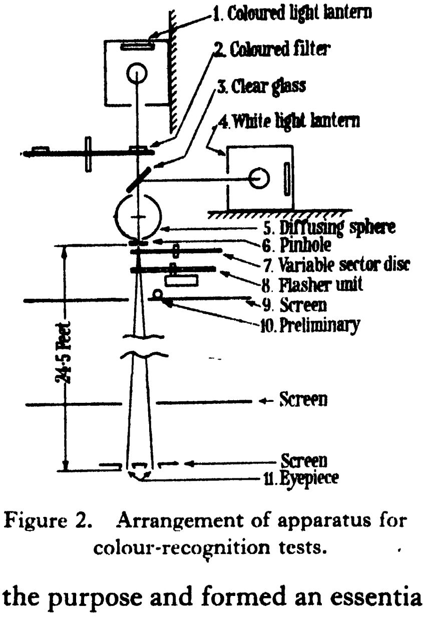


Figure 2. Arrangement of apparatus for colour-recognition tests.

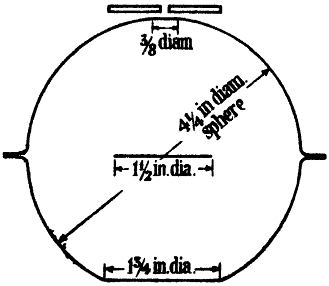


Figure 3. Diffusing sphere.

The tests were made in a photometric dark room, the black walls of which formed the general background of vision. Screens were placed as shown in figure 2 to intercept stray light from the test apparatus.

§ 5. CALIBRATION OF APPARATUS

Wratten light filters were used, consisting of 2-inch squares of coloured gelatine sandwiched between two clear-glass plates. During the tests there was no spectrophotometer available, and the properties of the filters were therefore calculated from the wavelength-transmission data published by the Kodak Co. (1938) and, in addition, were measured by visual photometry.

The total transmissions of the filters were measured by a flicker photometer, and the values were checked by comparison with standard filters on a Lummer-Brodhun contrast head. The colour coefficients were measured by means of a

Donaldson colorimeter. All these measurements were made by several observers whose results were averaged.

More recently the wavelength-transmission characteristics of the filters were measured by means of a photoelectric spectrophotometer, with the exception of No. 22, which had been so measured for threshold experiments soon after the conclusion of the present tests, and No. 23, which was no longer available. The measured spectral transmission values are given in table 1, together with certain additional data from the Kodak specification. From these measured values the total transmission and the colour coefficients were calculated for each filter for a 2848° K. source.

Table 2 gives a comparison of the filter properties obtained (*a*) by calculation from the Kodak specification, (*b*) by calculation from the spectrophotometric measurements, and (*c*) from the visual photometry. It will be seen that a very fair measure of agreement on colour exists among the three sets of data and, for most of the filters, the x, y coordinates of each set lie within ±0·005 of the mean. The chief exceptions are filters 34, 47 and 63, whose precise colours are therefore in doubt.

In view of this general agreement, and in spite of the lapse of time, it seems reasonable to assume that the measured spectrophotometric data are a fair representation of the filters at the time of the tests. The x, y coefficients given under (*b*) in table 2 have therefore been used to plot the test results on the colour diagrams except in the case of filters 23, 33, 45 and 73. For these four filters the coefficients given under (*a*) have been used because they agree more closely with the visual measurements under (*c*). For the sake of completeness the Kodak data for these four filters are included in table. 1

In the absence of other data the transmission values obtained by visual photometry were used to calculate the lantern adjustments, etc., for each particular signal (see § 7), but in any case the visual data, which were obtained by careful measurements, may be regarded as the most reliable assessment of the transmission values at the time of the recognition tests.

The 500-watt and 60-watt filament lamps, used as light sources in the coloured and white light lanterns respectively, were calibrated for 2848° K. colour temperature by matching with an N.P.L. standard lamp on a Lummer-Brodhun head. The lamps for the transmission and colour measurements were similarly calibrated.

The scale of each lantern was calibrated for white light in terms of the illumination at the observer's eyes. To do this the pinhole was removed and the brightness of the output window of the diffusing sphere was measured, for various positions of the lantern, by means of an illumination photometer. It was calculated, from the diameter of the pinhole and the distance of the observer's eyes, that an eye illumination of 1 mile-candle would require a brightness of 1·60 candles per square foot in the sphere window; thus a calibration curve of eye illumination against lantern position was obtained. The illumination photometer was itself calibrated by an N.P.L. standard candle-power lamp on a photometric bench.

Both the colour temperature and the illumination scale of the lamps were checked from time to time during the course of the tests. The diffusing sphere was twice cleaned and re-coated with magnesium oxide, and it was also found necessary to renew the silvered-glass mirror in the coloured-light lantern.

Table 1. Measured spectral transmission data for Wratten filters

K, data from Kodak specification. R, measured in 1940. Other data measured in 1946. Transmission in % at each wavelength.

λ (mμ)	15	22 R	23 K	24	31	32	33	33 K	34	45	45 K	46	47	47*a*	63	65*a*	71*a*	73	73 K	74
400	—	—	—	—	14·5	35·5	1·5	1·8	60·6	—	—	3·3	22·3	9·7	—	—	—	—	—	—
410	—	—	—	—	16·5	35·3	1·6	1·3	64·7	—	—	2·0	27·2	21·5	—	—	—	—	—	—
420	—	—	—	—	19·2	37·7	1·8	1·4	66·0	—	—	3·2	34·7	30·5	—	—	—	—	—	—
430	—	—	—	—	25·0	42·3	2·3	2·2	63·3	1·2	2·7	12·2	41·0	37·0	—	2·5	—	—	—	—
440	—	—	—	—	41·7	50·2	4·8	10·0	58·3	10·7	18·2	16·7	46·0	39·5	—	8·7	—	—	—	—
450	—	—	—	—	52·1	60·5	14·3	15·8	50·0	28·5	27·6	29·0	49·2	38·2	—	17·5	—	—	—	—
460	—	—	—	—	45·7	61·7	16·8	10·0	37·2	41·0	34·7	36·0	48·3	34·0	—	27·5	—	—	—	—
470	—	—	—	—	27·6	52·5	5·5	3·1	23·6	46·4	39·9	37·0	42·5	27·0	1·3	37·5	—	—	—	—
480	—	—	—	—	11·3	37·7	—	0·1	11·2	47·8	41·5	31·7	35·0	19·7	3·7	46·5	—	—	—	—
490	—	—	—	—	3·7	22·5	—	—	3·3	45·3	39·9	23·4	26·5	12·0	7·5	53·0	—	—	—	—
500	—	—	—	—	0·8	11·2	—	—	0·5	39·3	34·9	14·5	18·5	5·6	11·2	53·5	—	—	—	—
510	0·5	—	—	—	—	4·0	—	—	—	29·5	28·3	6·3	10·7	1·6	14·0	49·5	—	—	—	—
520	15·5	—	—	—	—	1·0	—	—	—	17·2	16·9	1·5	4·3	—	14·8	39·8	—	—	—	13·2
530	57·0	—	—	—	—	0·5	—	—	—	7·0	8·0	0·3	1·5	—	12·8	27·5	—	—	—	10·0
540	79·5	—	—	—	—	0·5	—	—	—	2·0	2·4	—	0·5	—	9·5	15·3	—	—	—	8·7
550	87·2	0·5	—	—	—	0·5	—	—	—	0·5	0·1	—	—	—	5·5	7·5	—	0·2	—	4·1
560	89·7	28·2	—	—	—	0·5	—	—	—	—	—	—	—	—	2·3	2·5	—	7·3	2·5	1·3
570	90·5	70·5	2·5	—	—	0·5	—	—	—	—	—	—	—	—	0·5	0·8	—	16·2	8·0	0·2
580	90·5	83·0	34·7	2·0	—	0·5	—	—	—	—	—	—	—	—	—	0·2	—	11·0	5·7	—
590	90·5	85·7	66·5	32·5	3·6	0·5	—	—	—	—	—	—	0·2	—	—	—	—	4·9	2·7	—
600	90·5	86·8	76·0	72·4	42·0	12·7	—	—	—	—	—	—	1·8	—	—	—	—	2·2	1·2	—
610	90·5	87·0	79·8	83·5	75·7	57·7	2·5	3·0	—	—	—	—	2·4	—	—	—	0·5	1·0	0·4	—
620	90·5	87·3	82·0	86·6	86·0	80·3	36·0	39·6	—	—	—	—	1·7	—	—	—	4·2	0·5	0·2	—
630	90·5	87·6	83·6	87·5	88·7	86·0	70·0	67·5	0·2	—	—	—	1·2	—	—	—	6·6	0·2	—	—
640	90·5	87·9	85·0	87·6	89·8	87·5	82·7	80·0	4·0	—	—	—	0·8	—	—	—	7·1	—	—	—
650	90·5	88·2	86·2	87·8	90·0	88·0	86·5	82·5	20·0	—	—	—	0·5	—	—	—	6·8	—	—	—
660	90·5	88·4	87·0	87·9	90·0	88·3	88·0	84·5	44·0	—	—	—	0·5	—	—	—	6·2	—	—	—
670	90·5	88·7	87·5	88·1	90·0	88·3	88·0	85·5	63·5	—	—	—	0·5	—	—	—	5·5	—	—	—
680	90·5	89·0	87·7	88·2	90·0	88·3	88·0	86·5	75·0	—	—	—	0·5	—	—	0·9	5·0	0·8	0·2	—
690	90·5	89·3	88·0	88·4	90·0	88·3	88·0	86·8	81·0	—	—	—	0·5	—	—	4·3	5·4	2·8	2·2	—
700	90·5	89·6	88·0	88·5	90·0	88·3	88·0	87·0	84·5	—	—	—	0·5	—	—	6·5	6·4	5·8	6·3	—

Table 2. Colour and transmission of Wratten filters with 2848° K. source

Wratten filter number	(*a*) Kodak specification			(*b*) Spectrophotometry			(*c*) Visual photometry		
	Colour		Transmission (%)	Colour		Transmission (%)	Colour		Transmission (%)
	x	y		x	y		x	y	
15	0·548	0·452	74·9	0·542	0·454	79·0	0·546	0·450	77·3
22	0·623	0·377	46·2	0·615	0·385	49·7	0·620	0·380	49·1
23	0·654	0·346	32·5	—	—	—	0·653	0·347	36·7
24	0·675	0·325	25·0	0·675	0·325	26·8	0·673	0·321	26·6
31	0·609	0·263	19·4	0·606	0·267	22·0	0·607	0·272	20·9
32	0·543	0·241	20·2	0·553	0·243	19·0	0·551	0·240	19·7
33	0·675	0·267	9·90	0·686	0·254	10·0	0·675	0·269	9·6
34	0·253	0·064	0·97	0·316	0·095	2·00	0·353	0·115	1·90
45	0·105	0·258	4·00	0·104	0·250	4·13	0·112	0·258	3·3
46	0·121	0·109	1·06	0·122	0·109	1·24	0·123	0·110	1·05
47	0·138	0·070	1·13	0·170	0·118	2·32	0·184	0·131	2·1
47*a*	0·145	0·043	0·46	0·142	0·052	0·65	0·134	0·058	0·52
63	0·120	0·670	3·54	0·159	0·689	3·71	0·152	0·693	3·1
65*a*	0·110	0·437	7·12	0·118	0·436	9·28	0·121	0·436	8·0
71*a*	0·709	0·291	1·29	0·710	0·290	0·90	0·712	0·288	1·31
73	0·496	0·503	1·87	0·490	0·510	3·93	0·500	0·497	4·25
74	0·200	0·766	2·62	0·225	0·750	1·92	0·232	0·745	1·9

§6. CONDITIONS OF EXPERIMENT

When giving the results of photometric, colorimetric and other tests involving visual observation, it is desirable to state the precise conditions under which the observations were made. The conditions under which the recognition measurements were made are therefore summarized here.

Two series of recognition tests were conducted, the first at an eye-illumination of 1 mile-candle and the second at an eye-illumination of 2 mile-candles. The point sources were viewed by binocular foveal vision, with dark-adapted eyes, against a dark background for a period of 1½ seconds. The angular diameter of the point source was 0·6 minutes of arc. The general background brightness was about 0·0001 candles per square foot, or of the order of brightness of a starlit sky.

The various colours were shown in succession in random sequence; there was an approximately equal number of each class of colour so that no class was unduly emphasized. During the tests each colour was seen alone and could not be contrasted with any other light. The tests were performed by nine male observers who were considered, from the results of certain transmission measurements and colorimeter tests, to have normal colour vision. The observers were tested by the Ishihara colour charts and all classed as normal. The age groups of the observers are shown in table 3.

Table 3. Age groups of observers

Age:	20–24	25–29	30–34	35–40	>40	Average: 30
Number:	3	2	2	1	1	Total: 9

§7. EXPERIMENTAL PROCEDURE

Seventeen coloured filters were used, and from these, by the addition of white, a total of 73 colours was produced. Each colour was identified by the number of the Wratten filter followed by a letter representing the degree of relative saturation of the colour as compared with the pure filter colour. The settings of the two lanterns and the adjustment of the variable sector disc were calculated for each colour. These settings, together with the identification number and letter of the colour, were written down on small index cards, there being one card to each colour and eye-illumination. The two series of 1 mile-candle and 2 mile-candles were taken separately and each series was divided into two groups to avoid tiring the observers. The cards in a group were shuffled before each test to preserve a random sequence.

The procedure in carrying out a test was as follows. The observer was allowed to get dark-adapted (usually about 10 minutes was sufficient for this), then, having adjusted his eyepiece, he observed each colour in turn. The observer was required to say, after each flash, what colour he considered the signal to be, the choice of colour being restricted to the following five categories: red, yellow or orange, white, green, blue. A definite decision was required on each observation, and no repetition of a colour was permitted unless the observer failed to see the signal because he blinked or was out of position. After each observation the operator readjusted the apparatus in accordance with the settings indicated on the next card. An assistant kept the voltage on the lamps at the correct settings, and also recorded the identification number and letter from each card together with the colour category named by the observer. The tests were repeated from day to day, over a period of nine months, until each colour had been seen by each observer 20 times, and in the case of the first group 30 times. The alternative designation of yellow or orange was permitted because some observers experienced a psychological reluctance to be limited to yellow. In assessing the results all yellow and orange designations were taken together.

§ 8. EXPERIMENTAL RESULTS

The observations recorded during the tests were sorted and tabulated for each colour. Table 4 shows this tabulation in the case of colour 31G. The successive readings of each observer are given, together with his total number of recognitions in each colour category. The results are added for all the observers and the percentage recognition evaluated. The symbol Y in the table includes both yellow and orange designations.

Table 4

Recognition record for colour	31 G
Eye illumination	2 mile-candles
Filter saturation	0·6

Observer	Recognition		R	Y	W	G	B
L. N. B.	R R R R Y R Y R Y R	R R Y R R R Y R Y Y	13	7	–	–	–
E. S. C.	Y R Y R Y W R R R W	W W W R Y W W Y Y Y	6	7	7	–	–
J. C. C.	R R Y R Y R R R R R	Y R R R R Y R Y Y R	14	6	–	–	–
S. H. G. C.	Y Y Y Y Y Y Y Y Y Y	Y Y Y Y Y Y Y Y Y Y	–	20	–	–	–
H. N. G.	Y R Y Y W Y Y Y Y R	Y R R Y R Y R R R R	9	10	1	–	–
N. E. G. H.	R Y Y Y R Y Y R R Y	Y Y R R Y R R Y R R	10	10	–	–	–
R. E. L.	R R R R R R R R R Y	R R R R R R R R R R	19	1	–	–	–
D. B. McK.	R R R R R R R R R R	R R R R R R R R R R	20	–	–	–	–
J. W. S.	R R R R Y R R R R R	R Y R R Y R R R R R	17	3	–	–	–
		Total	108	64	8	0	0
		Percentage	60	36	4	0	0

The percentage recognitions thus found were plotted as ordinates against the relative saturation, S, as abscissae, one set of curves being plotted for each filter at each eye illumination. Figure 4 shows the set of curves for filter No. 45 at 2 mile-candles illumination.

Interpolation on the above curves enabled recognition contours to be plotted on the colour diagram. In a few cases the highest recognition required a relative saturation greater than unity, and extrapolation of the curves was made in these cases; the resulting points on the contours lie between the pure filter points and the spectrum locus.

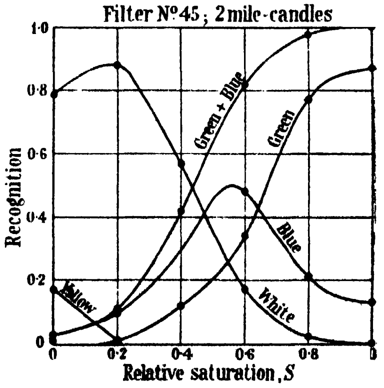


Figure 4. Typical saturation-recognition curves.

The method of calculating the colour coefficients corresponding to a given relative saturation of a filter is a particular case of the general problem of calculating colorimetric purity which has been analysed by Judd (1931). It can be shown that, referring to figure 1, if x_w, y_w, z_w are the coefficients of the white point W, x, y, z, are those of C, and x_f, y_f, z_f, those of F, then

$$x = P \, . \, x_f + (1 - P)x_w \qquad \ldots\ldots(1)$$

and

$$y = P \, . \, y_f/S, \qquad \ldots\ldots(2)$$

where

$$1/P = 1 + (1/S - 1)y_f/y_w, \qquad \ldots\ldots(3)$$

and where S is the relative saturation of the filter as defined in § 7. The quantity P, given by equation (3), is the ratio of the distance CW to the distance FW, and may be termed the *relative excitation purity*.

§ 9. DISCUSSION OF RESULTS

The tests which have been described produced a total of 30,420 observations taken over a period of nine months. The results have been summarized in the form of colour-recognition contours plotted on the standard I.C.I. colour diagram in figures 5 and 6, for 1 and 2 mile-candles, respectively.

The data should be strictly applied to signals having the same spectral-energy distribution as those used for the experiments, but it seems improbable that recognition can be critically dependent on spectral-energy distribution, and such isolated rough checks as have been possible suggest that a wide variation of spectral distribution has a relatively small effect on recognition. It therefore seems reasonable to apply the present results to the coloured signals which are used in practice, but further experimental data are required to confirm the validity of this procedure.

It was found that green and blue signals could not be distinguished from one another with any certainty at the low illuminations used for the tests. It is clear that green and blue would not be suitable for use as two separate signal colours at long range. The recognition values of green and blue were therefore added to obtain recognition contours for a single signal colour called green + blue in figures 5 and 6. This procedure does not imply any new restriction on the choice of signal colours in practice, as blue signals (i.e. signals which appear blue at short range) are in any case unsatisfactory at long range because the eye has difficulty in focusing them and because, owing to the low luminosity of blue light, blue filters have low transmission values.

A noticeable feature of the results is that the points plotted in figures 5 and 6 readily form smooth contours for the green + blue and for the white, but are not satisfactory for the red group. In the case of the yellow + orange group the points were so scattered that it was thought best to draw the nearest smooth curve through them. The importance to be attached to the yellow + orange contours is therefore considerably less than to the contours of the other colours, and this is emphasized by the fact that the highest yellow recognition point available was 80% at 2 mile-candles and 70% at 1 mile-candle, as compared with 100% for green + blue and red. It is concluded that yellow + orange is the least satisfactory group for a signal colour at low illumination.

In view of the poor recognition of the yellow + orange group it would be expected that it would be very difficult indeed to obtain orange as a separate group. This was in fact the experience during the tests.

Lines representing the limits defined in B.S.S.563/1937 (see Appendix) for aviation colours have been drawn. It will be seen that the areas thus defined are areas of high recognition except in the case of white, which extends into the region of yellow recognition. It appears that the specification for aviation white was framed so as to include the paraffin flame, but there seems to be little justification for this and, in view of the danger of confusion with yellow, it is thought that the specification should be amended so that, instead of "x not greater than 0·540", it should read "x not greater than 0·477". Filament lamps operating at colour temperatures down to 2500° K., including all the lamps usually used for aviation purposes, would thus fall within the specification.

If it be assumed that 80% or higher recognition is satisfactory, there is a considerable area of satisfactory green + blue recognition outside the B.S. specification. This additional area is, however, not a useful area because of the practical objections to blue signals already mentioned. There is also a large area of high red recognition outside the specification, but if the specification were extended to cover this area, which lies in the blue direction, the short-range appearance of the colours might be unsatisfactory. It is clear that any extension of the red specification along the spectrum in the direction of shorter wave-lengths would be unsafe.

The data presented in the contours of figures 5 and 6 were obtained under conditions where atmospheric absorption has no appreciable effect on the results, and where no searching of the field of view was required. No precise data are yet available on the change of colour of light transmitted through hazy atmosphere, or on the influence on recognition of searching the field for the signal. The comparison with the B.S. specification for aviation colours has not therefore taken into account these two factors.

The effect of increasing the illumination at the eye from 1 to 2 mile-candles is not anywhere very great. The greatest effect is noticed in the green + blue region, where the recognition is raised about 10%. The deep blue, white (on the purple side), and red (except spectrum reds) are unaffected. It seems likely that further increase of illumination would not give a proportionate increase of recognition except possibly in the case of the yellow + orange group. It also appears that small errors in the adjustment of the intensity of the signals will not have had an important effect on the results.

All the purples used in the tests appeared red when seen as point sources, the usual dichromatic characteristic of purple point sources being absent.

Tests were made to determine whether a flash period longer than 1½ seconds would affect the results. No change in recognition could be detected. The recorded results were studied closely to discover whether the results of test depended on the order in which the observer viewed the colours. No such dependence could be detected, and it was concluded that any given observation was unaffected by the previous observation.

It would naturally be of the greatest interest to compare the data of the present paper with the results obtained by other investigators. There are two other sets of published results which were obtained by somewhat similar test methods and which might be taken for comparison. The recognition tests of McNicholas have already been referred to. Another series of tests was made at about the same time as the present tests and subsequently published by Holmes (1941).

The results obtained by McNicholas in his tests unfortunately suffer from two

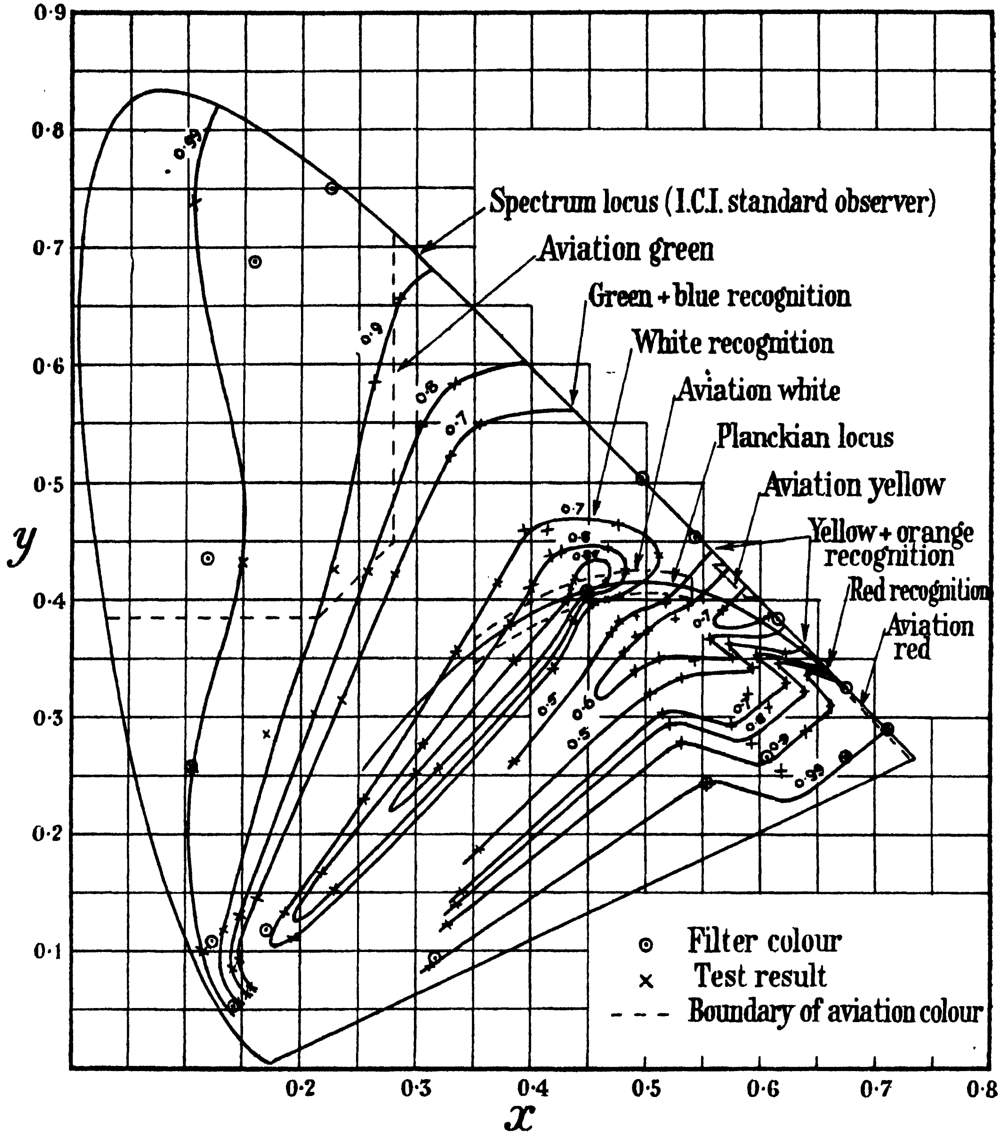


Figure 5. Colour-recognition contours for 1 mile-candle point sources.

limitations: first, the results represent the average recognition of signals whose illuminations range from 0·40 to 6·2 mile-candles in one series of tests, and between still wider limits for other series; and second, only a single line is given for each colour category instead of an area of recognition as in the present tests. It is therefore difficult to compare the results of the present tests with McNicholas's results. It is, however, of interest to note that he finds that green and blue are not easily distinguished, a conclusion which is in agreement with the results of the present tests.

Holmes's tests were made with an apparatus of excellent design, and had the advantage that 256 coloured signals were used. Unfortunately these tests also suffer from a severe limitation; it is that each signal was seen no more than three times by each of six observers. Since most of the signals were likely to be recognized, at least occasionally, in three or more of the colour categories, it is clear that three observations per observer are quite insufficient to yield statistically significant results. The large number of signals used does not compensate for this deficiency. It is therefore not surprising that, while a general similarity with the present results exists, there is no detailed concordance.

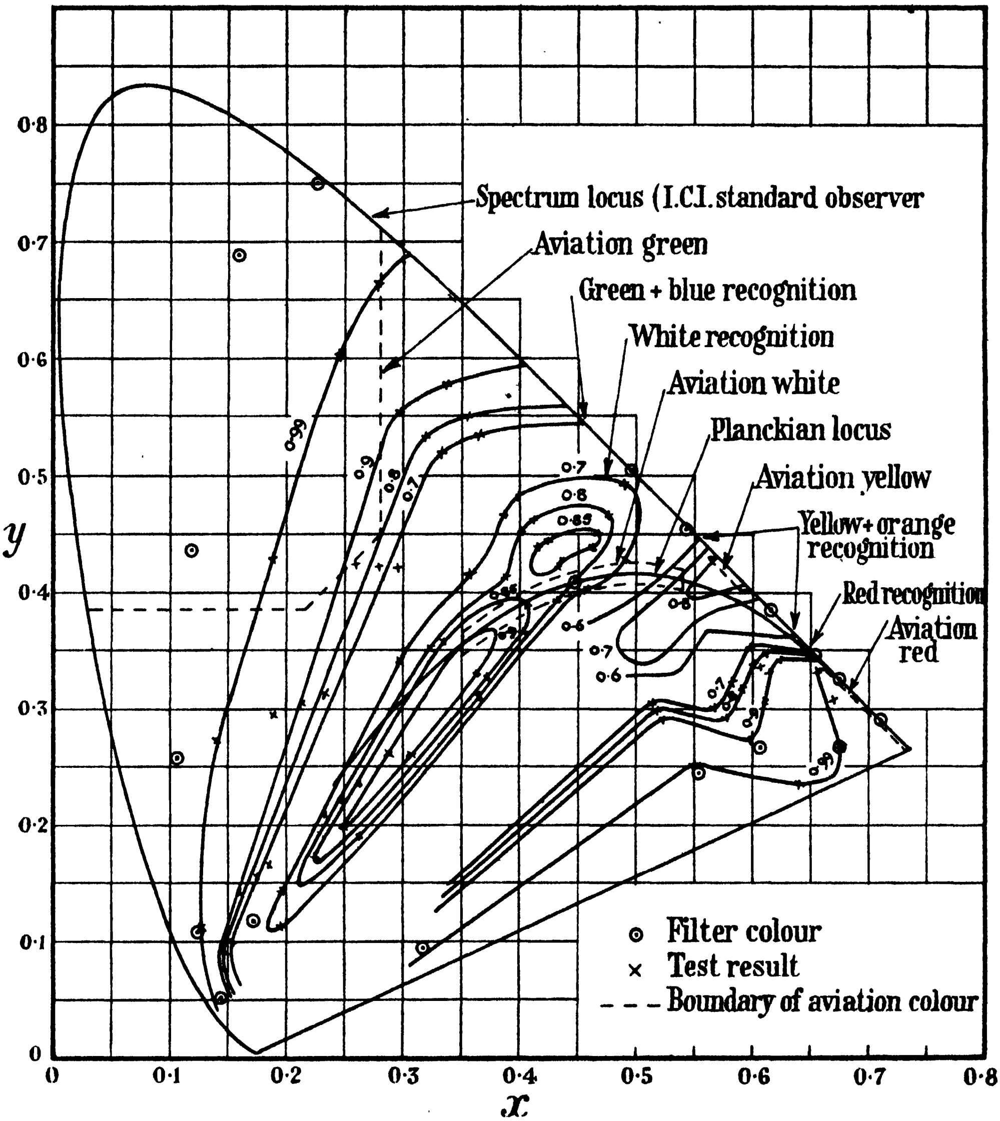


Figure 6. Colour recognition contours for 2 mile-candles point sources.

§ 10. ACKNOWLEDGMENTS

The author wishes to express his gratitude to colleagues who acted as observers during the tests. He also wishes to thank the Chief Scientist, Ministry of Supply, for permission to publish this paper.

APPENDIX

Aviation colours (B.S.S.563/1937. Appendix A)

	x	y	z
Aviation red	—	$\leqslant 0{\cdot}335$	$\leqslant 0{\cdot}002$
Aviation yellow	—	$0{\cdot}402$ to $0{\cdot}430$	$\leqslant 0{\cdot}007$
Aviation green	$\left.\begin{matrix}\leqslant 0{\cdot}280 \\ \leqslant y-0{\cdot}170\end{matrix}\right\}$	$\geqslant 0{\cdot}385$	—
Aviation white	$0{\cdot}350$ to $0{\cdot}540$*	—	—

* $|x-y_0| < 0{\cdot}01$, where y_0 is the y-coordinate of the Planckian radiator for which $x_0=x$.

REFERENCES

Guild, J., 1931. *Phil. Trans. Roy. Soc.*, A, **230**, 149.
Holmes, J. G., 1941. *Trans. Illum. Engng. Soc.*, **6**, 71.
Judd, D. B., 1930. *Bur. Stand. J. Res.*, **4**, 515.
Judd, D. B., 1931. *Bur. Stand. J. Res.*, **7**, 827.
McNicholas, H. J., 1936. *Bur. Stand. J. Res.*, **17**, 955.
Stiles, W. S., Bennett, M. G. and Green, H. N., 1937. *A.R.C. Technical Report R. & M.* No. 1793.
Wratten Light Filters, 1938. Eastman Kodak Co.

THE MEASUREMENT OF THE CHROMATIC AND ACHROMATIC THRESHOLDS OF COLOURED POINT SOURCES AGAINST A WHITE BACKGROUND

By N. E. G. HILL,
Royal Aircraft Establishment, Farnborough, Hants

MS. received 24 *September* 1946

ABSTRACT. Measurements were made during 1939–40 to determine the effect of background brightness on the recognition of aviation light signals. White, yellow, red, and green point-source singals were observed by monocular foveal vision against a white background whose brightness was varied from 10^{-3} to $2{\cdot}6\times10^{1}$ candles/sq. ft., a range of brightness from less than that of a starlit sky to that of a clear noon sky 20° from the sun. From the results of repeated observations of these signals curves were drawn showing the chromatic and achromatic thresholds and also the photochromatic ratio of the four colours as functions of background brightness. The curves were drawn for 50% recognition, and it is estimated that the thresholds for reasonable certainty of recognition are from three to five times those given. It is concluded that yellow is a comparatively unsatisfactory colour at both very low and very high background brightnesses.

§ 1. INTRODUCTION

It has been found that, when observing coloured light signals which are near the limit of visibility, the minimum signal intensity at which it is possible to recognize the colour of the signal is, in general, higher than the minimum intensity at which it is possible to detect the presence of the signal. That is to say, if the intensity of a signal be progressively reduced, the colour of the signal will disappear before the signal is lost to view. The intensities at which the colour of a signal ceases to be recognizable, and at which the signal ceases to be visible, are known as the chromatic and achromatic thresholds respectively, and the ratio of these intensities is called the photochromatic ratio of the signal. The threshold values and the photochromatic ratio are functions of the brightness of the background against which the signal is observed.

The threshold intensities of light signals are not sharply defined values, below which the signal is never seen and above which it is always seen. There is a range of intensities over which the signal will sometimes be recognized, sometimes be seen but not recognized, and sometimes not be seen at all. There are thus several ways of defining the threshold values, and we might, for instance, define the achromatic threshold either as the intensity below which the signal will never be seen, or as the intensity above which the signal will always be seen. Unfortunately these definitions, admirable in theory, do not lead to specific values in practice, and it is therefore more convenient to define the achromatic threshold at a particular background brightness as the intensity which will make the signal visible on an average of 50% of the occasions on which observation is attempted. Similarly, we shall take the chromatic threshold as the intensity at which the colour of the signal will be correctly recognized on an average of 50% of the occasions on which observation is attempted.

Some data are already available on chromatic and achromatic thresholds of monochromatic visible radiations against a black background, but only for test lights of appreciable angular size. These data have been summarized by Stiles, Bennett and Green (1937). In the absence of any data for point-source signals, tests were made at the Royal Aircraft Establishment during 1939–40 to determine the thresholds of aviation light signals. The results of these tests, which could not previously be published owing to wartime restrictions, are given in the present paper.

When dealing with point-source signals it is convenient to refer to the illumination at the observer's eye rather than to the intensity of the signal. All thresholds are therefore given as values of eye illumination in mile-candles.

§ 2. METHOD OF TEST

The determination of the threshold values was made by repeated observations of a series of signals, in a manner similar to that used by the author to measure the colour-recognition values of coloured light signals (1947).

The method was to fix the background at a particular brightness and then to present a succession of signals in random sequence to an observer who was required to place each signal in one of a number of colour categories or, if he failed to observe the signal, in the category " nil ". The signals were of four colours, white, yellow, red, and green, and there were signals of several values of eye-illumination for each colour. The signals were repeated many times and observed by a number of different observers. The average recognitions of the various colours and of "nil" were plotted, and from the curves the chromatic and achromatic thresholds for the particular background brightness were obtained.

The tests were repeated at various values of background brightness from complete darkness to a value of brightness equivalent to that of the clear noon sky about 20° from the sun. Care was taken to maintain the background at the same white colour throughout the series of tests.

§ 3. DESCRIPTION OF APPARATUS

It was necessary to arrange for point-source signals to be observed in the centre of a bright background, and to ensure that the colour and intensity of the point source could be changed rapidly and accurately so that a regular succession of signals could be seen by the observer. A smooth presentation of the signals greatly relieves the observational strain in this type of test and thus leads to more reliable results.

(a) *Optical arrangement*

The general arrangement is shown diagrammatically in figure 1. The "point source" consisted of a 24-volt, 36-watt filament lamp, with a compact coiled-coil filament, mounted in a matt black screening box. Two such point sources were fitted, one for direct vision and the other to be seen by reflection at 45° in a clear glass optical flat; a ten-to-one ratio of intensity was thus obtained. A further range of intensities was obtained by means of a stepped variable sector disc giving eight values of transmission from 0·5 to 100%. Either of the filament lamps could be exposed to view by means of solenoid-operated shutters, and three coloured filters were provided so that white, yellow, red or green signals could be produced. The filament lamps were selected so that the maximum dimension of the light source, including bulb reflection, was not more than 0·1 inch.

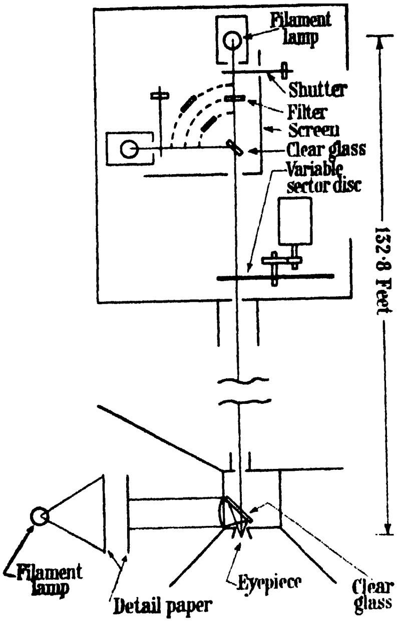


Figure 1. Arrangement of apparatus for threshold measurements.

The point-source signal was viewed by monocular vision by an observer seated 132·8 feet away. The point source therefore subtended an angle not greater than 0·25 minutes of arc. The observer's eye was located by means of a rubber eyepiece, but was not restricted by an artificial pupil. The background, which was superimposed on the signal by reflection from a clear glass optical flat, consisted of two sheets of detail paper spaced about 2 inches apart and illuminated from behind by a filament lamp. The background was viewed through a condenser lens placed so that the observer's eye was at the focus of the lens, and the observer saw an image of the bright detail paper at infinity. With this arrangement there was no difficulty in focusing the eye on the point-source signal. The bright field, which was bounded by the periphery of the lens, was circular and subtended 45° at the observer's eye.

As a help to the observer, and in order to stabilize the test results, four fixation points (not shown in figure 1) were placed at the corners of a square, of side subtending 1°·5 at the eye, and the signal appeared at the centre of the square. The fixation points were just bright enough to be seen with certainty above the background brightness. The whole apparatus was carefully screened to prevent stray light disturbing the observations.

(b) *Operational arrangements*

In order to secure rapid setting of the coloured signal, the variable sector disc was mounted on a sliding carriage whose position was varied by means of a D.C. motor controlled by a relay circuit and a set of position-selecting switches. The coloured filters were mounted in pivoted holders which could be swung in front of either of the signal filament lamps by moving three-position levers. Each lever, in addition to placing the corresponding filter in position, also closed a contact in series with the appropriate shutter solenoid.

The sector-disc and filter settings having been made, the signal was presented to the observer through a timing circuit, controlled by an electrically maintained pendulum of 1-second period. The pendulum contacts operated on a 3-second cycle in such a way that, when the push button had been pressed, a single-stroke gong warned the observer, the shutter opened for about 2½ seconds, then closed, and finally the relays were returned to rest ready for the next signal. It was found that a new signal setting could be prepared within 3 seconds, so that a continuous series of signals at 6-second intervals could be presented to the observer.

The tests were carried out with the arrangements described above at a series of background brightnesses obtained by using various sizes of filament lamp at various distances from the detail paper. In certain cases minor modifications were necessary. At the highest brightness the observer was moved to a distance of 40·1 feet from the signal source in order to obtain sufficient eye illumination from the signal. At this distance the source subtended 0·7 minutes of arc. The fixation points were suitably spaced to remain on a 1°·5 square. The brightest background was obtained by replacing the detail paper with ground glass on which was projected a defocused image of a projector lamp filament. A second condenser lens was used to flash the field of view. At the low background brightnesses it was necessary to reduce the signal intensity, and this was done by interposing a diffusing sphere in front of the direct-viewed filament lamp. In front of the sphere was placed a 0·1-inch diameter aperture.

§4. CALIBRATION OF APPARATUS

All filament lamps used during the tests, both for the signals and for the backgrounds, were calibrated for 2848° K. colour by matching with an N.P.L. standard colour-temperature lamp, and were operated throughout at that colour temperature.

The candle power of each signal lamp at 2848° K. colour was measured by standard visual photometry using the Lummer-Brodhun contrast head. In the case of the diffusing sphere arrangement, the candle power was measured using a 0·407-inch diameter aperture and the corresponding value with the smaller aperture calculated. The values of eye illumination of each signal were then calculated.

The background brightness was measured at each setting using a portable brightness photometer to transfer the brightness to the standard photometer bench.

The spectral transmission curves of the coloured filters were measured on a photoelectric spectrophotometer, and the curves are given in figure 2. The colour coordinates and total light transmission of the filters, in conjunction with a 2848° K. colour source, were calculated and are given in table 1.

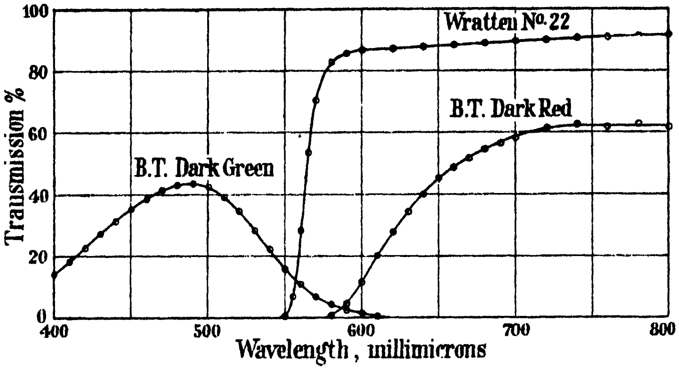


Figure 2. Spectral transmission curves of signal filters.

The white, red, and green colours are within the limits specified for aviation colours given in B.S.S. 563/1937. The yellow colour is somewhat more orange than aviation yellow.

Table 1. Colour coordinates and light transmission of coloured filters with 2848° K. colour source

Filter	Colour coordinates		Transmission (%)
	x	y	
B.T. dark green, N.P.L. 102/1924	0·184	0·392	11·90
B.T. dark red, N.P.L. 102/1924	0·693	0·307	8·18
Wratten No. 22	0·615	0·385	49·7

Colour of background is 2848° K. colour temperature.

§ 5. CONDITIONS OF TEST

The visual conditions of the tests may be summarized as follows.

Point-source signals were viewed by monocular foveal vision for about 2½ seconds against a circular white background subtending 45° at the observer's eye. The angular diameter of the point source was not more than 0·25 minutes of arc except in the case of the brightest background, when it was not more than 0·7 minutes of arc. The observer's pupil was unrestricted, but fixation points were used.

White, yellow, red, and green coloured signals of various eye illuminations were shown in succession in random order, and the background brightness was varied in steps from approximately 10^{-5} candles/ft² to 2610 candles/ft².

Observations were made by eight male observers of normal colour vision. The age groups of these observers are given in table 2.

Table 2. Age groups of observers

Age :	20–24	25–29	30–34	35–40	>40	Average : 32
Number :	1	3	2	1	1	Total : 8.

§ 6. TEST PROCEDURE AND RESULTS

About five values of eye illumination were chosen for each colour, giving a total of about twenty signals for each background brightness. The settings of the apparatus for each signal were written on a small index card and the cards were shuffled to obtain a random sequence. The signals were then presented to the observer successively in the sequence given by the cards, and the observer's response to each signal was written on the corresponding card.

The observer was given a short period to become adapted to the background. In the case of the dark background, a period of 10 minutes was allowed for dark adaptation. The signals were then presented successively at 6-second intervals, and the group was repeated four times, so that about 100 signals were seen at a sitting, which occupied about 10 minutes. Short rest intervals were permitted when required by the observers. The tests were repeated at other sittings until each observer has seen each signal 25 times. The nine background brightnesses used involved a total of about 35,000 observations.

The test observations were grouped as follows :—White, yellow (including orange), red, green (including blue), nil.

The results of the eight observers were classified, added, and the percentage recognition of each of the above groups calculated. These recognition percentages are plotted as ordinates against eye illumination, *E* in mile-candles, as abscissae in figures 3, 4, 5, and 6 for signals which were actually white, yellow, red, and

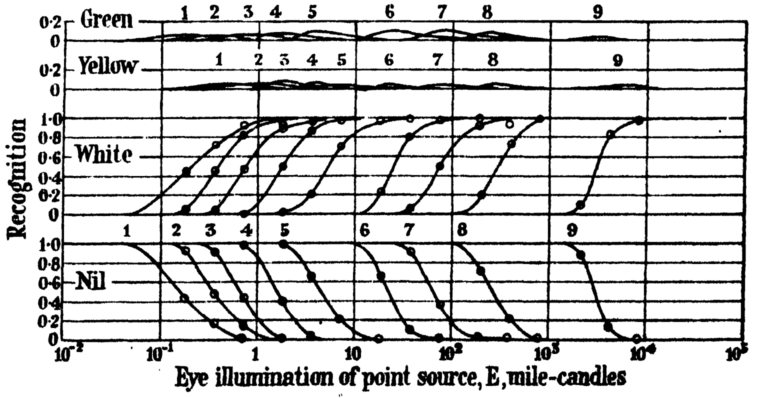


Figure 3. Recognition curves for white point source. Curve numbers refer to background brightnesses in table 3.

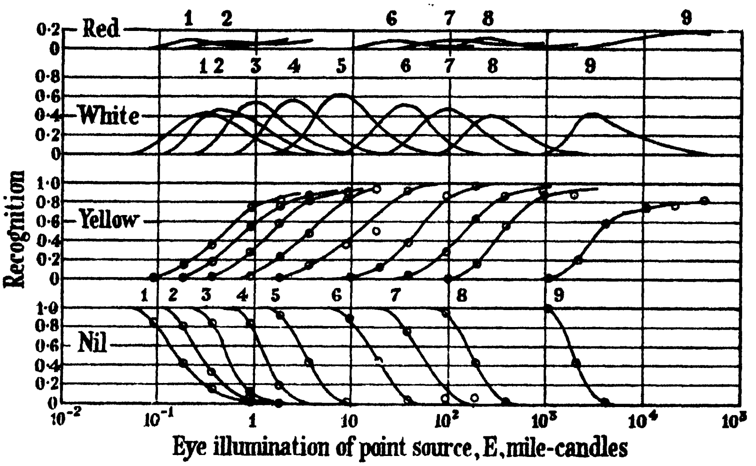


Figure 4. Recognition curves for yellow point source. Curve numbers refer to background brightnesses in table 3.

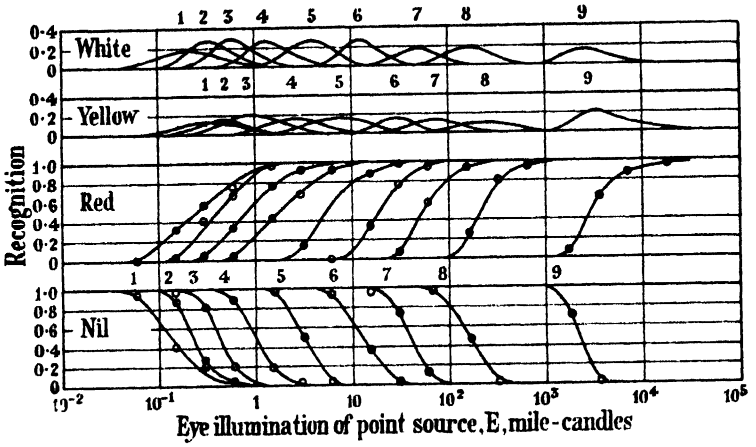


Figure 5. Recognition curves for red point source. Curve numbers refer to background brightnesses in table 3.

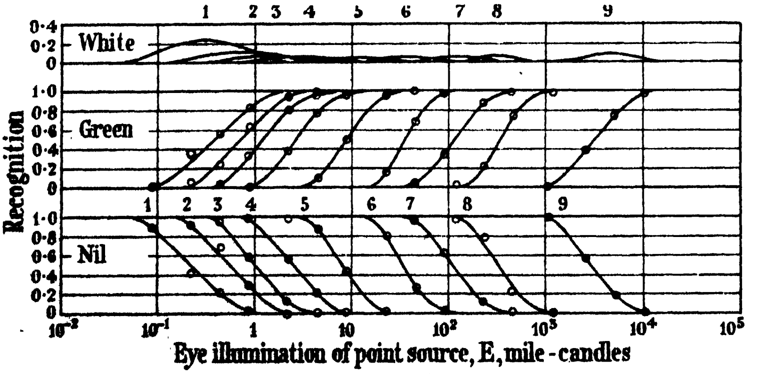


Figure 6. Recognition curves for green point source. Curve numbers refer to background brightnesses in table 3.

green respectively. The figures show families of curves, each curve at a constant background brightness, B, whose value is given in candles per square foot in table 3.

Table 3. Background brightnesses corresponding to curve numbers in figures 3–6

Curve number	Brightness, B (candles/ft^2)	Log$_{10}$ B
1	Approx. 10^{-5}	$\bar{5}$·0
2	0·0111	$\bar{2}$·05
3	0·0508	$\bar{2}$·71
4	0·298	$\bar{1}$·47
5	1·75	0·24
6	10·3	1·01
7	47·1	1·67
8	292	2·47
9	2610	3·42

The achromatic threshold values for white signals can now be found by considering the "nil" recognition curves in figure 3, for evidently the 50% chance of detecting a signal is the same as the 50% chance of not seeing it. Hence the 50% "nil" ordinate gives the threshold value of eye illumination of the signal corresponding to each value of background brightness. Similarly the chromatic threshold for each value of background is obtained by reading off the values of illumination corresponding to 50% recognition of the true colour of the signal, in this case white, in figure 3.

It is clear from the general form of the "white" and the "nil" curves in figure 3 that threshold values based on the certainty of seeing or recognizing the signal, or on the certainty of not seeing or recognizing it, cannot be obtained with any reasonable precision. The reason for choosing the 50% recognition criterion for threshold values is thus apparent. It is however possible, and indeed of some interest, to find the achromatic and chromatic thresholds for 10% and 90% recognition; these values can be obtained from figure 3 fairly satisfactorily, bearing in mind that the 10% and 90% achromatic thresholds correspond to 90% and 10% recognition of "nil" respectively. In figure 7 the threshold values of illumination of white signals are plotted as functions of background brightness for 10%, 50% and 90% recognition.

In a similar manner, the thresholds for yellow, red, and green point sources are shown in figures 8, 9 and 10, the values being obtained from figures 4, 5 and 6 respectively.

The relation between the achromatic thresholds of the four signal colours is shown in figure 11, and between the chromatic thresholds in figure 12, for 50% recognition.

The photochromatic ratio, p, was calculated by taking the ratio of the 50% chromatic to the 50% achromatic threshold for each colour, and values of $\log p$ are plotted against values of $\log B$ in figure 13. In effect, figure 13 represents the ratio of figures 11 and 12.

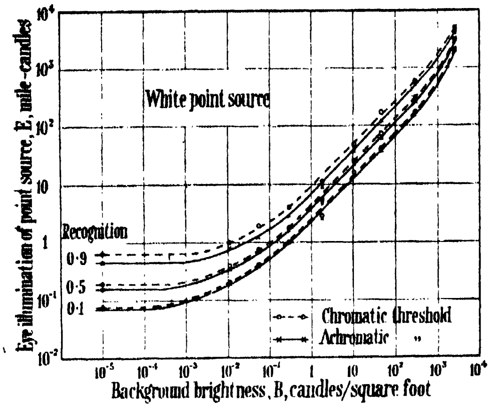


Figure 7. Threshold values of white point source.

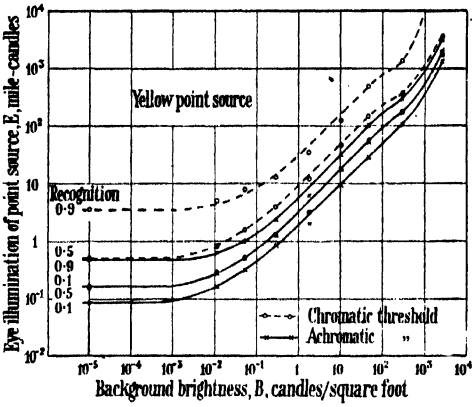


Figure 8. Threshold values of yellow point source.

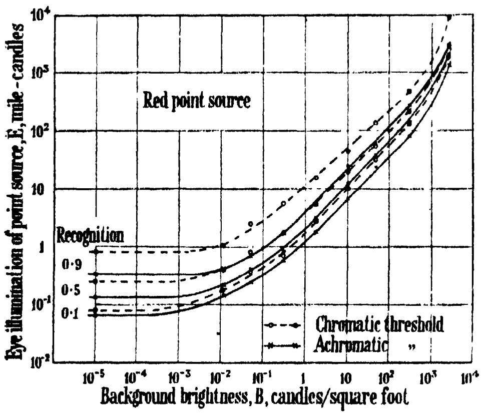


Figure 9. Threshold values of red point source.

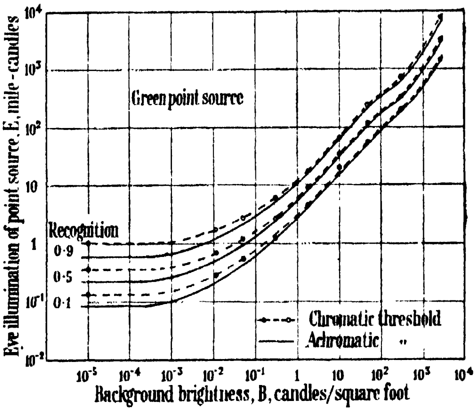


Figure 10. Threshold values of green point source.

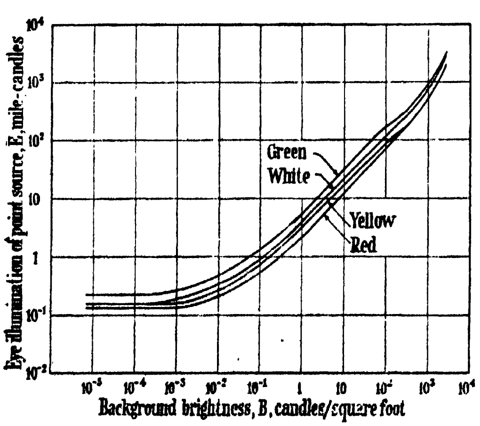


Figure 11. **Achromatic threshold values for 0·5 recognition.**

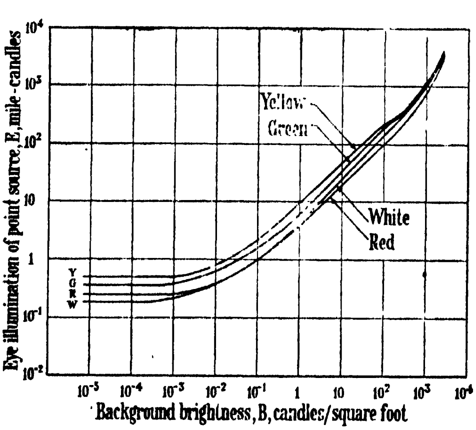


Figure 12. **Chromatic threshold values for 0·5 recognition.**

§ 7. DISCUSSION OF RESULTS

The achromatic and chromatic threshold curves given in figures 11 and 12 enable a prediction to be made as to whether a signal of a particular colour and eye illumination will be visible against a given background. Alternatively, for given background conditions, it is possible to estimate the illumination required to make a signal visible or recognizable, and hence the intensity of signal required to cover a particular range. It must be remembered, however, that the curves in figures 11 and 12 are drawn for threshold values at which there is an even chance that either the signal will or will not be observed in the case of the achromatic threshold, or that the signal will or will not be recognized correctly for colour in the case of the chromatic threshold. It will be seen from figures 7–10 that there is a considerable range of uncertainty both for achromatic and chromatic recognition, and that the uncertainty range is in general greater at the very low backgrounds. The 90% achromatic threshold varies from about 3 to about 1·5 times the 50% threshold with increasing background brightness, and the 90% chromatic threshold from about 3 times to about twice the 50% threshold except in the case of yellow, whose 90% threshold is much higher. It thus appears that, for certainty of observation, a signal would need to be less above the threshold at high background brightnesses than at low ones. This, however, is not the case in practice, because the data given here were obtained under observational conditions which did not require the observer to search his field of view for the signal, whereas, under normal conditions in aviation, the observer does not know precisely where to look. Furthermore the eye is assisted in its search at low background brightness by the extra foveal sensitivity of the retina when the eye is dark adapted (i.e. the background less than 10^{-3} candles/sq. ft.). Thus, under practical conditions, the uncertainty thresholds are likely to be from 3 to 5 times the thresholds given in figures 11 and 12.

The values of recognition for each colour group form families of related curves in figures 3–6, and the experimental points fall very well on to the individual curves. There is, therefore, an indication that the results are self-consistent, and also that a sufficiently large number of observations was made to yield statistically satisfactory averages.

The curves show the extent to which colour confusion occurs when the illumination of the signal is below the certainty level. Thus in figure 3 the white point source receives a certain amount of green and yellow recognition, but never more than 10% of either colour. The red point source in figure 5 sometimes receives as much as 30% white recognition and sometimes as much as 20% yellow recognition, although the two colour confusions do not occur together. The green point source in figure 6 may have nearly 35% white recognition against the dark background, but this confusion falls to less than 10% as the background brightness is raised. There is no indication of any confusion whatever between green and red for either the red or the green signals.

The curves in figure 4, for the yellow point source, exhibit rather different characteristics from those for the other three coloured signals. When the signal illumination is below the chromatic threshold, the recognition of the signal as white may be more than 60% at the medium background brightnesses, a value very much greater than for any of the other there signal colours. The red recognition reaches values of 10%, much the same as in the case of white signal, but, unlike that case, the red recognition curve, having reached a maximum, falls and then rises again to a second maximum at an illumination corresponding to the highest yellow recognition.

The important feature of this particular yellow signal (Wratten No. 22) is therefore that, at both low and high background brightness, its recognition as yellow fails to reach 100%, even when the illumination is well above the chromatic threshold, because of confusion with red. The author's previous work on colour recognition (1946) showed that yellow is a comparatively unsatisfactory signal colour for point sources against a dark background, and that the particular yellow now under discussion was likely to be confused with red. The present results confirm this view, and also show that this yellow is equally unsatisfactory against very bright backgrounds.

Figures 11 and 12 reveal a curious bend in the curves for the green and yellow signals at a background brightness of about 500 candles/square foot; there is no trace of any similar effect with the white or red signals. The results of some threshold measurements on a green signal made some years before the present tests, using an extinction method, suggest that the valve of background brightness at which the bend occurs is a characteristic of the individual observer. It therefore seems likely that the occurrence of the bends in the curves for the green and the yellow signals is caused by certain properties of the retina. A very interesting theory of rod-and-cone sensitivity has been put forward by Stiles (1939) which may provide an explanation, but, owing to the complexity of the theory, it has not so far been possible to apply it to the present data.

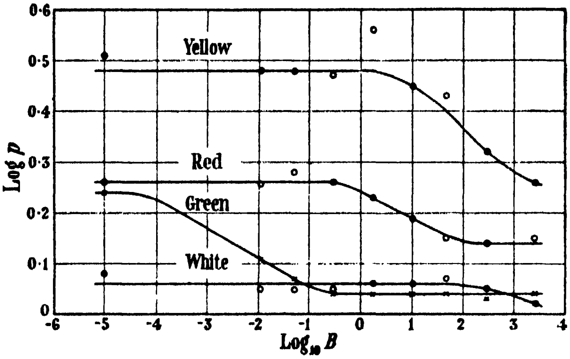


Figure 13. Photochromatic ratio for 0·5 recognition.

The photochromatic ratio curves for 50% recognition, shown in figure 13, exhibit certain interesting features. In spite of the fact that the scale for log p is rather extended, the points are found to lie very closely on smooth curves. The ratio, p, for a white signal is little greater than unity, and the ratio for green is also near unity except for very dark backgrounds, when the ratio rises to about 1·7. The value of p for red is about 1·8 for low backgrounds and falls to 1·4 for high backgrounds; this is not entirely in agreement with the common experience that a red signal looks red to extinction, but it is probably due to red-yellow confusion lowering the red recognition. The value of p for yellow is about 3 at low backgrounds and continues level until at bright backgrounds the ratio falls sharply. The existing data on the photochromatic ratio are scanty, but such data as exist do not appear to contradict the results embodied in figure 13.

Certain data exist for achromatic thresholds for dark-adapted foveal vision, and the mean values derived from these data by Stiles, Bennett and Green are: white, 0·24; red, 0·14; green, 0·32 mile-candles. Referring to figure 11, the achromatic thresholds for dark backgrounds are: white and yellow, 0·16; red, 0·14;

green, 0·22 mile-candles. Thus the present tests give the same value for red, but lower values for white and green thresholds, although the white and green thresholds are in the same ratio in each case. It is possible that a certain amount of extra-foveal recognition has occurred, lowering the white and green values but leaving the red unaffected.

§ 8. ACKNOWLEDGMENT

The author wishes to thank the Chief Scientist, Ministry of Supply, for permission to publish this paper.

REFERENCES

Hill, N. E. G., 1947. *Proc. Phys. Soc.*, **59**, 560.
Stiles, W. S., 1939. *Proc. Roy. Soc.*, B, **127**, 64.
Stiles, W. S., Bennett, M. G. and Green, H. N., 1937. *A.R.C. Technical Report R. & M.* No. 1793.

A TIME MICROMETER OF HIGH ACCURACY

By E. A. NEUMANN,
Scophony Research Laboratories, Wells, Somerset

MS. received 12 *November* 1946

ABSTRACT. A water-ethyl alcohol mixture having a zero temperature coefficient of ultrasonic velocity over a certain temperature range having been discovered, the development of an accurate time micrometer using such a mixture was attempted but was found to meet with difficulties due to partial evaporation tending to alter the composition of the liquid. Further research, however, led to the discovery that at an elevated yet convenient temperature—of 72°·7 c.—water itself displays a zero temperature coefficient of ultrasonic velocity over a useful range. A time micrometer using distilled water was therefore developed.

§ 1. INTRODUCTION

Scophony Ltd. in pre-war days developed their television receiving system which was based on the Debye-Sears effect of light diffraction by ultrasonic waves in a liquid (Scophony Ltd. and Jeffree, 1934). Experiments were carried out in this connexion for the purpose of discovering a liquid in which the speed of the ultrasonic waves would be constant over a reasonable range of temperatures. It had been found that the temperature coefficient of ultrasonic velocity in water displayed an anomalous behaviour; whereas in other pure liquids the velocity fell with rising temperature, in the case of water it increased. Efforts were therefore made to find a mixture of water and some suitable liquid in which the temperature coefficients would just compensate each other to give a zero temperature coefficient. Ethyl alcohol, well known to be miscible in all proportions with water, was tried and found suitable (Scophony Ltd. and Jerram, 1940), a mixture of ethyl alcohol and distilled water containing 16% of alcohol having a zero temperature coefficient of ultrasonic velocity at temperatures at and near 20° c.

It was quickly realized that the possibility of controlling the temperature dependence of ultrasonic velocity opened up a wider field of applications for ultrasonic waves than the one for which this possibility had originally been sought. Thus, Scophony Ltd. and Dodington (1940) suggested the use of an ultrasonic cell as a frequency stabilizer in an oscillator circuit. Again, on the suggestion of A. F. H. Thomson, formerly of the Scophony research staff, the Ministry of Supply approached the company with the suggestion that they should develop an instrument for the very accurate measurement of short time intervals, using ultrasonic waves travelling over a variable and accurately measurable distance at a known velocity kept constant within exceedingly close limits. The accuracy required was, in the course of the work, specified as *ca.* $\frac{1}{20}$ microsecond at any part of the scale, which was to be calibrated from 5 to 240 microseconds.

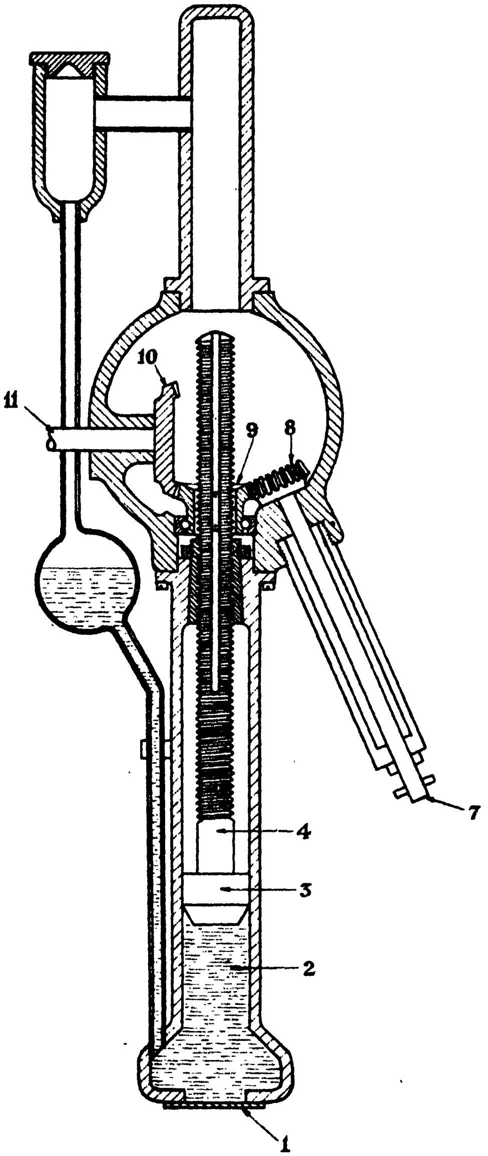


Figure 1. Sectional sketch of time micrometer.

§ 2. GENERAL DESCRIPTION

The instrument which was eventually developed, and which was of the same general type as that initially envisaged, is illustrated in figure 1. Here 1 is a piezo-electric crystal having electrodes in the form of metal coatings and which, when driven by a pulse from a suitable oscillator, will vibrate and thereby generate a train of waves in the liquid 2. This train of waves, after travelling towards a plane reflector constituted by the surface of steel plunger 3, is reflected by it and returned to the crystal 1 where it produces a second pulse. The two pulses are suitably amplified and made visible on the screen of a cathode-ray tube; plunger 3 is moved by means of micrometer screw 4 rotated via gear wheels 8 and 9 by a manually operated shaft 7, until the distance between the two deflections on the cathode-ray tube screen due to the two pulses is the same as the distance corresponding to the time interval which it is desired to measure. This time interval will thus be related to a length on the micrometer scale (which, in the device shown, is connected to the micrometer screw 4 by means of another gear wheel 10 engaging gear wheel 9 and operating the spindle 11 of a suitable indicating device not shown). It will be seen that this method of measurement partakes of the advantages of a compensation method (such as a measurement on a Wheatstone bridge) in that any non-linearities in the time base of the cathode-ray tube are ineffective, the time measurement being reduced to the *equalization* of, instead of ordinary comparison between, two distances on the cathode-ray tube screen.

The remaining parts of the instrument illustrated in figure 1 are self-explanatory.

§ 3. DEVELOPMENT AND TESTS OF THE INSTRUMENT

The chief task in developing the instrument consisted (*a*) in a thorough investigation into the way in which the accuracy of measurements is affected by temperature fluctuations and into means to overcome the difficulty so caused, and (*b*) in a sufficiently precise determination of the ultrasonic velocity under operating conditions.

For this purpose, measuring apparatus was developed comprising two piezo-electric crystals of a standardized frequency of 18 Mc./sec., one of which was fixed near one end of a trough in which ultrasonic waves were to be produced, while the other crystal was mounted on a carriage movable along through the liquid contained in the trough. Both crystals were provided with metal coatings acting as electrodes, the first one acting as a transmitter of ultrasonic waves, for which the second acted as a receiver. The transmitting crystal was driven from a 100 Kc./sec. temperature-controlled quartz bar oscillator, the ouput of which underwent a number of stages of frequency multiplication to arrive at the required 18 Mc./sec. The output of the oscillator, in addition, underwent frequency division down to 50 c./sec., which frequency was used to drive a domestic clock, and by comparing the readings of this with the Greenwich time signals, the crystal driving frequency could be checked with abundant accuracy. The output of the receiving crystal was passed through an amplifier specially designed to ensure that its output voltage amplitude was constant and that the phase of this voltage remained fixed relative to the phase of its input. This output voltage, together with a portion of the voltage driving the transmitting crystal, was applied to a valve phase comparator, the output of which fed a meter and counter. If the movable carrier with the receiving crystal was moved towards the transmitting crystal, the needle of the meter fluctuated over almost the whole scale as the phase of the waves at the face of the receiving crystal relative to that of the waves leaving the surface of the transmitting crystal changed through 2π. The counter was arranged to increase its reading by one unit per 2π period. The receiving crystal carriage was fitted with a stop consisting of an insulated micrometer head. An 18-inch length standard bar calibrated by the National Physical Laboratory was used, its face nearest the transmitting crystal resting against a stop provided at that end. The receiver crystal carriage was driven automatically towards the other end of the standard bar, the counter operating each time the receiving crystal face advanced by one whole wave-length, until electrical contact with the standard bar was established, when the carriage was immediately brought to rest by the driving clutch being disengaged. The micrometer stop was then adjusted, moving the carriage a very short distance further towards the standard bar until the next operation of the counter occurred. The counter and micrometer were then read, after which the standard bar was swung clear of the carriage, which thereupon proceeded to advance at a speed of ½ mm. per second, until it touched the stop near the transmitting crystal end, against which the standard bar had previously rested. The micrometer was then adjusted again until the counter operated, whereby it was ensured that a whole number of wave-lengths had been traversed in the run of the carriage. This number was obtained by subtracting the first from the second counter reading. The corresponding distance traversed was obtained as the sum of 18 inches and the difference in the micrometer readings, the measurement thus supplying all the data required for determining the ultrasonic velocity; if the distance traversed is l, the number of waves in it N, and the frequency f, then the veolcity, ${}_TV^c$, with indices T and c indicating its dependence on temperature and alcohol concentration, is equal to ${}_TV^c = l/N.f$.

In the experiments carried out, l, N and f were all determined to an accuracy better than 1 part in 10,000. The temperature of the liquid was kept constant by circulating thermostatically controlled water through the double walls of the trough provided for this purpose. Several precision thermometers were immersed in the liquid, and frequent checks of the alcohol concentration were carried out gravimetrically. This latter point, however, proved one of the main difficulties attending both the preliminary experiments and the proposed design of the actual instrument, as ethyl alcohol, as is well known, evaporates at a considerably higher speed than water, and the strength of the mixture was thus strongly inclined to alter. Thus it was found that at *ca.* 25° C. the ultrasonic velocity in a mixture containing approximately 16% of alcohol changed by *ca.* 2½% in 18 hours.

In order, therefore, to arrive at a reliable figure for the ultrasonic velocity, several series of experiments were required, and were carried out, as follows:—

I. So as to arrive at a correction for partial evaporation, the trough was closed and sealed, a mixture comparatively rich in alcohol (over 20%) was placed in the trough and its concentration gradually and continuously reduced by the addition of water. Samples of the mixture were abstracted from the trough at intervals and its constitution determined gravimetrically. The changes in number (and fractions) of waves were continuously read from the phase comparator.

II. The actual velocity measurement had to be carried out with the trough open, as it was not otherwise possible to move the carrier. The concentration of the water-alcohol mixture was gravimetrically determined at the instants beginning and ending each run, and a correction derived from a curve representing the results of the measurements carried out under I was applied to each number of waves, as follows:—If T and c are, as before, temperature and alcohol concentration of the mixture, l is the accurately known length of approximately 18 inches as explained before, L is the distance between the crystals at their near position, ${}_TN^c_L$, ${}_TN^c_{L+l}$ etc. are the numbers of wave-lengths at temperatures, concentrations and distances indicated by the several indices, and ${}_TN_m$ the measured number of wave-lengths, then

$$\begin{aligned} {}_TN_m &= {}_TN^c_{L+l} - {}_TN^{c-\Delta c}_L \\ &= {}_TN^c_{L+l} - {}_TN^c_L + \int_c^{c-\Delta c} \frac{d({}_TN^c_L)}{dc}\, dc, \end{aligned}$$

whence

$$ {}_T N^c_{L+l} - {}_T N^c_L = {}_T N^c_L = {}_T N_m - \int_c^{c-\Delta c} \frac{d({}_T N^c_L)}{dc}\, dc. $$

Measurements were carried out at an approximately constant temperature of 25° c., a small correction (of 0·0016 inch) being applied to the length of the standard bar, which had been calibrated at 62° F. = 16°·5 c. or 8°·5 below the measuring temperature. The final result of this series of measurements, with corrections applied, is shown in figure 2, which shows that the required accuracy was obtained, deviations from the mean straight line not exceeding 1 part in 10,000 represented by 0·16 metres/sec. in the velocity ordinate.

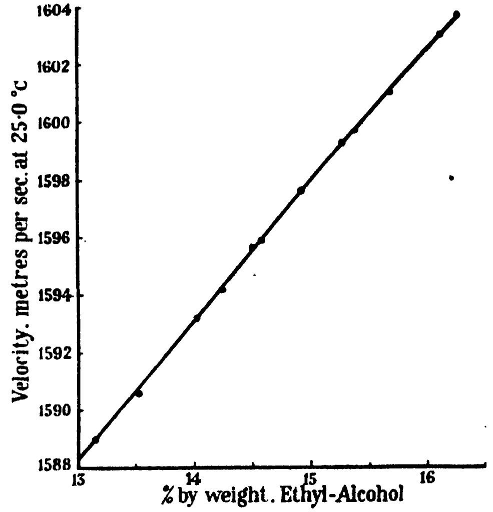


Figure 2. Velocity of ultrasonic waves in water-ethyl alcohol mixture, as dependent on concentration.

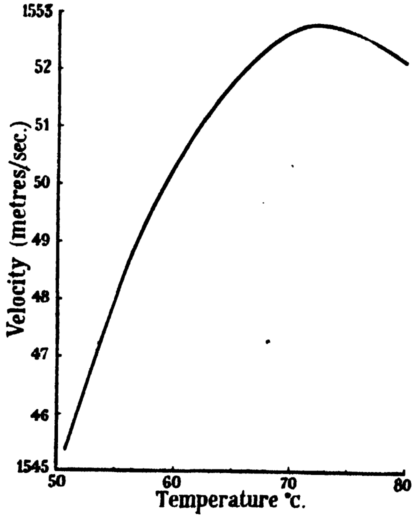


Figure 3. Dependence on temperature of the ultrasonic velocity in distilled water.

III. Measurements at different temperatures and different concentrations were also carried out with the trough closed and sealed. From these measurements a few facts emerged at once: (*a*) It was found that the attenuation of supersonic waves was very rapid below 20° c. (*b*) Different mixtures have a zero velocity-temperature coefficient at different temperatures. (*c*) No mixture was found for which the region of substantially zero velocity-temperature coefficient covered an extended temperature interval; the useful interval in fact proved to be substantially constant over the whole range investigated and to amount to 6° c. if a variation of not more than 1 in 10,000 was permitted.

Because of (*a*) and (*c*), it was decided to embody some temperature control in the final instrument to keep the temperature at a somewhat elevated point. There remained the difficulty of the mixture being inclined to change its composition, due to the different rates of evaporation of its components.

To overcome this difficulty, it was suggested either to use a mixture of which only the water component was liable to evaporate to any sensible degree, e.g. sodium iodide/water or glycerine/water, and periodically to "top up" the mixture with water, or to use a mixture both components of which evaporate at or near the same rate, e.g. to mix water with propyl alcohol (the vapour-pressure curve of which is almost identical with that of water). The latter method was favoured.

There followed, however, a discovery which, besides being interesting in itself, considerably simplified the problem. This was that pure water behaved in a way similar to that in which the water-ethyl alcohol mixtures had been shown to behave: the ultrasonic velocity in it as plotted against its temperature went through a maximum.

The following method was now adopted for arriving at a curve accurately relating ultrasonic velocities to temperatures. The receiving crystal was placed about 25 inches from the transmitting crystal and fixed, and the trough, filled with distilled water of 0·6 megohms per cm^3 at 20° c., was covered. The temperature was raised to *ca.* 80° c. and allowed to fall slowly. As, in consequence, the supersonic velocity changed, the number of wave-lengths between the two crystals also changed, leading to a gradual relative change of phase between input and output voltage, whole multiples of 2π of which were registered on the counter. The temperature was read each time the counter operated until 51° c. was reached. The difference in the counter reading, ΔR_1, from that at 51° c. was recorded against temperature. The whole procedure was repeated with the receiving crystal the exact length of the 18-inch standard bar nearer to the transmitting crystal, and the difference of the counter reading from that at 50° c., ΔR_2, also recorded against temperature. In evaluating the results, it had to be borne in mind that, although the counter recorded phase change in multiples of 2π, it gave no indication of whether the phase was advancing or retarding at these instants. This could, however, be determined by ascertaining whether the meter needle was moving in the same or the reverse direction, as when the carriage was given a slight movement towards the transmitting crystal; it was in this way known whether the number of wave-lengths was increasing or decreasing at any temperature. With this knowledge it was possible to ascertain ΔN_1 and ΔN_2, the difference in the number of wave-lengths in the distance separating the crystals at the temperature T, and at 51° c. and 50° c. respectively, for the two separations 25 inches and 7 inches approximately. It was found from curves relating ΔN_1 or ΔN_2 to T that the maximum of ultrasonic velocity in distilled water of the stated degree of purity occurs at about 72°·7 c. At this temperature a complete run over substantially the length of the standard bar was taken, as explained earlier in this article, to arrive at the ultrasonic velocity at this particular temperature, which was found to be equal to 1552·7 metres per second. Velocities at other temperatures were now derived from the measurement leading to the value at 72°·7 c. and from the curves relating ΔN_1 and ΔN_2 to T, thus saving a considerable amount of time as compared with that which would have been needed to measure the velocities at various temperatures in the same way as at 72°·7 c.

Let x be the exact length of the 18-inch bar, and the several symbols and indices having the meanings explained hereinbefore, then

$$\Delta N_1 = {}_{51}N_{L+x} - {}_{T}N_{L+x},$$

$$\Delta N_2 = {}_{50}N_L - {}_{T}N_L,$$

and

$${}_{T}N_x = {}_{T}N_{L+x} - {}_{T}N_L = ({}_{51}N_{L+x} - {}_{50}N_L) - (\Delta N_1 - \Delta N_2).$$

${}_{7\cdot 27}N_x$ was measured when the ultrasonic velocity at 72°·7 C. was determined, and was found to amount to 5300·4 wave-lengths, and, at the same temperature, $\Delta N_1 - \Delta N_2$ was found to amount to 25·6. It follows from this that

$$ {}_TN_x = 5326{\cdot}0 - (\Delta N_1 - \Delta N_2), $$

and, therefore,

$$ V_T = \frac{fx}{5326{\cdot}0 - (\Delta N_1 - \Delta N_2)} \text{ inches per second} $$

$$ = \frac{8229{\cdot}6 \times 10^3}{5326{\cdot}0 - (\Delta N_1 - \Delta N_2)} \text{ metres per second.} $$

V_T is shown plotted against T in figure 3 (Jones and Gale, 1946).

After this discovery had been made, it was decided to operate the time micrometer with a distilled-water filling and at *ca.* 73° C., and the instrument was constructed accordingly. The thermo-controls of the micrometer were so devised that on starting operations a primary or auxiliary heater was put into action which quickly raised the temperature of the whole to a temperature in the vicinity of the correct operating value of 73° C.; after 30 minutes, the supply to the auxiliary heater was automatically switched off by one of the bimetallic switches incorporated in the device, the temperature being subsequently maintained solely by the maintenance heater. The correct operating temperature of 73° C. was attained after approximately another ten to fifteen minutes.

The traversal of the plunger (3 in figure 1) was operated by gearing driven by a hand-wheel; gear ratios 1:1 and 4:1 could be obtained by pulling out and pushing in the hand-wheel. Provision was made for automatically disengaging the drive from the plunger at the two ends of the range over which it was to operate, and for re-engaging it on reversal of the hand-wheel.

The time range to which the range of plunger travel was to correspond had been specified by the users as being from 5 to 240 microseconds. Keeping in mind that the distance from crystal to plunger surface was travelled over twice by the ultrasonic waves (once before and once after reflection), this corresponds to a distance from crystal to plunger of from approximately 4 mm. to approximately 192 mm.

The counter was so calibrated that one unit on it corresponded to about 1·2 microseconds (the exact figure had been specified by the users), and that $\frac{1}{20}$ of a unit could be read on any part of the scale. By the side of the counter was a thermometer, underneath a water-level indicator, near the top of the front panel of the apparatus a green "tell-tale" lamp indicating correct operating conditions, and near the lower edge terminals and power switch. A push-button served as an automatic cut-out and reset switch which operated the primary heater.

The internal mechanism of the micrometer was stainless steel throughout to prevent corrosion.

ACKNOWLEDGMENTS

Research preparatory to the design of the instrument was carried out by A. J. Gale, to whom several suggestions, including that to use pure water, are due, P. L. F. Jones being in general charge of the research and development. The

author, in writing the present article, has freely drawn on a report by the latter. The details of the actual design were dealt with by A. E. Adams.

Thanks are due to the Director of Scientific Research, Ministry of Supply, for permission to publish the results of this work.

REFERENCES

Jones, P. L. F. and Gale, A. J., 1946. *Nature, Lond.*, **157**, 341.
Scophony Ltd. and Dodington, S. H. M., 1940. Brit. Pat. No. 573,269.
Scophony Ltd. and Gale, A. J., 1947. Brit. Pat. No. 582,435.
Scophony Ltd. and Jeffree, J. H., 1934. Brit. Pat. No. 439,236.
Scophony Ltd. and Jerram, C. F., 1940. Brit. Pat. No. 534,448.
Scophony Ltd. and Thomson, A. F. H., 1947. Brit. Pat. No. 582,434.

COLORIMETRY IN THE GLASS INDUSTRY

By J. G. HOLMES,

A lecture given to the Colour Group 14 *November* 1945; *MS. received* 27 *September* 1946

ABSTRACT. Recent knowledge of the glassy state has enabled the theories of modern colour chemistry to be applied to glass, and developments in colorimetric technique have put the design and performance of coloured glasses on a quantitative basis. The methods of colour measurement particularly suited to transparent media are described, together with rapid approximate methods of calculation. The properties of the important colouring oxides are given, and the effects of concentration, thickness and illuminant are discussed. The colours and reflexion-factors of bloomed glass surfaces, the properties of some special colour filters and a basis for specification of coloured glasses are briefly described.

§ 1. INTRODUCTION

Although the glass industry is by no means a large user of colorimetric methods, this lecture must be restricted to a few of the applications of colorimetry and, therefore, to those with which the author is most familiar. Amongst the important items which are not discussed are the colours of decorative and domestic glass, coloured opal glass and similar surface colours, photoelectric colorimetry and the terminology of glass colours. The subjects discussed will include a very brief statement of the background knowledge of coloured glass, the methods of measurement and calculation appropriate to a transparent medium and one or two points of interest, including some special colour filters, the colours of "bloomed" lenses with non-reflecting films and the basis of specification for coloured light signals. This last item is the one to which colorimetric methods are most widely applied, as glasses obtained through ordinary commercial channels are not usually closely graded, and it was also one of the first industrial applications of the trichromatic system agreed in 1931 by the Commission Internationale de l'Eclairage.

§ 2. THE STRUCTURE OF GLASS

It is not enough to say that glass is a "super-cooled liquid". It has recently been defined by Scholes (1945) as "an inorganic product of fusion which has cooled to a rigid condition without crystallizing". It is typically hard and brittle, but it may be colourless or coloured, transparent or opaque. Its structure is similar to that of the liquid state characterized by so high a viscosity that it is for all practical purposes rigid (Morey, 1938). Just as liquids are analogous to crystals, so there may be a close similarity between the arrangements of atoms in a glass and in a crystal, even though there is no crystalline structure in glass.

The current theory of glass regards it as a three-dimensional network consisting mainly of silicon and oxygen atoms in random orientation, each silicon atom being bonded to four oxygen atoms and each oxygen atom to two of silicon, and, as the bond is very strong, the properties of silica glass are very stable. Other atoms, such as boron, which have glass-forming oxides, may take their place in the network, but mostly the other elements used in glass-making go into the holes in the network and will generally loosen the silicon-oxygen network and alter the physical properties. For example, the addition of sodium oxide to silica may be represented by sodium ions in holes adjacent to oxygen atoms which are bonded to only one silicon atom, causing amongst other things a lowering of the softening temperature and an increase in the thermal expansion coefficient. If boric oxide is added to the soda-silica glass, it forms part of the network, reducing the number of single-bonded oxygens and tightening up the whole system, reducing the thermal expansion coefficient. If lead oxide is introduced, the lead goes into the holes in the network and makes a heavier softer glass and, in general, elements of different atomic weights and different valencies and affinities will yield glasses of different properties. Some elements will give an unstable system of forces between the ions, and this instability is associated with selective absorption of light. The deepest absorption bands, which give the deepest colours when they occur in the visible region of the spectrum, are associated with ions of two different valencies of the same element, and, for example, manganese will normally give a rich purple colour, due to unstable balance between the oxidized and reduced states, but this colour is almost completely absent if the manganese is strongly reduced. Ferrous iron gives a strong absorption band in the infra-red and ferric iron gives a strong absorption band in the blue, but in practice it is difficult to obtain either complete reduction or complete oxidation, and glasses coloured with iron are subject to control of the ferroso-ferric balance.

The system of forces in the network will depend on both the composition of the network-formers and the modifiers or chromophore ions, and thus the absorption bands may be affected by the base glass as well as by the colouring ingredients. For example, cobalt usually gives a blue colour in soda-silica glasses, but it gives a reddish colour in borate or phosphate glasses which are highly acidic, and it gives a pink colour if it replaces the sodium in a soda-silica glass. Titanium has the property of modifying the network and loosening its structure, so that an ion which normally goes into a hole in the network may take up a silicon position in the network itself, and the colouring effect of the ion is greatly affected even though titanium produces no coloration itself. An example of this is a ceria-soda-silica glass which is colourless but which becomes a pronounced yellow when titania is introduced.

Raising the temperature of glass will loosen the structure and reduce the differences between the energy levels, so that the absorption bands tend to move towards longer wave-lengths which have a smaller quantum of energy.

This network-model of the glassy state is far from complete, but it provides a very useful basis for argument. It is developed in some detail in a monograph by Weyl (1944) now being published in the *Journal of the Society of Glass Technology*.

The difference between glass and crystals is illustrated by the absence of any sharp change in refractive index associated with an absorption band in coloured glass. There may be some relation between the rise in index and the rise in absorption towards the ultra-violet end of the transmission spectrum, where the absorption is due to the forces in the network rather than to a modifying ion, and if so, this would show the family relationship between the random network in glass and the regular network in crystals.

§ 3. THE PROPERTIES OF A TRANSPARENT COLOURED MEDIUM

Glass is an excellent example of a coloured material, because the effects of absorption under different circumstances can be calculated or estimated from comparatively simple data and there are no effects of texture or gloss. Colour is the subjective effect of selective absorption in the visible spectrum, and the colour name given to a glass is the complement of the colour which is absorbed. In fact, glass is a simple example of the subtractive process of producing colour, and it follows that subtractive instruments such as the Lovibond Tintometer are as suitable for measurement of coloured glasses as they are for coloured liquids.

Greater concentration of the colouring constituent in a glass, or greater thickness of glass, will give a lower transmission factor (the glass-maker's equivalent of lightness or brightness of surface colours) and will usually give a purer, more saturated colour. Considering a cobalt blue glass as an example, figure 1 shows the transmission-wave-length curves of six glasses containing increasing amounts of cobalt and figure 2 shows the colours of the light, from a source operated at a colour temperature of 2848° K., after transmission through each of the glasses. The strongest absorption band in figure 1 is at about 600 mμ, and so the colour of a pale cobalt glass in figure 2 is on the side of the source towards the complementary wave-length. As the amount of cobalt increases, the orange radiation near 600 mμ is almost completely absorbed and the absorption in the yellow-green region becomes more noticeable, so that the colour locus moves away from the yellow-green region. The darkest glass in figure 1 shows absorption of substantially all the orange-yellow-green radiation and the colour of the transmitted light may be represented by the centre of gravity of the wave-length bands from 400 mμ to about 500 mμ and from about 680 mμ to 750 mμ. In general, figure 2 shows how the hue and saturation of the transmitted light change as the amount of cobalt is increased. It may be noticed that although the process is subtractive, the author's thinking is in terms of additive mixing of the light which is transmitted.

Another important property of a glass is its transmission factor, which is the percentage ratio of the amount of light transmitted through the glass to the amount of light incident on it, evaluated in terms of the standard visibility function. The absorption of light necessary to produce colour means that coloured glasses

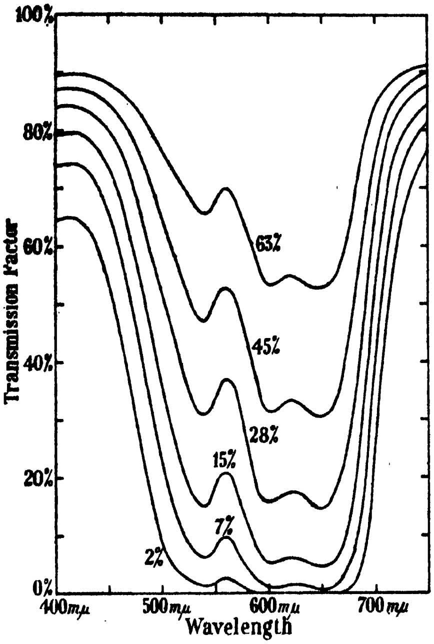


Figure 1. Transmission curves of six cobalt glasses. (The figures show percentage total transmission factor.)

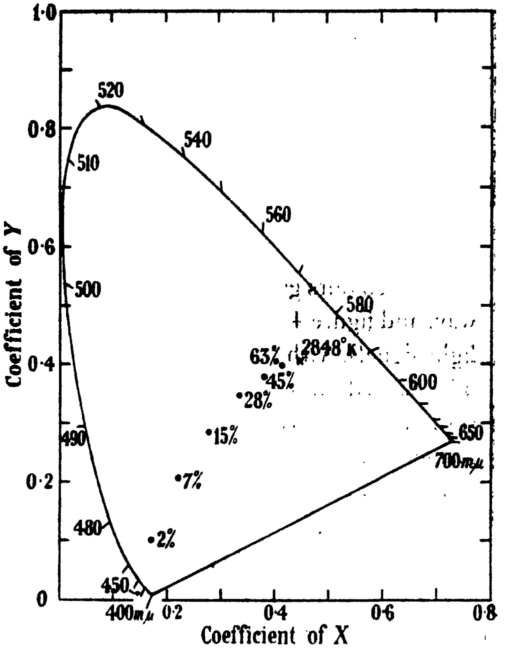


Figure 2. Colours and transmission factors of six cobalt glasses with 2848° K. light source.

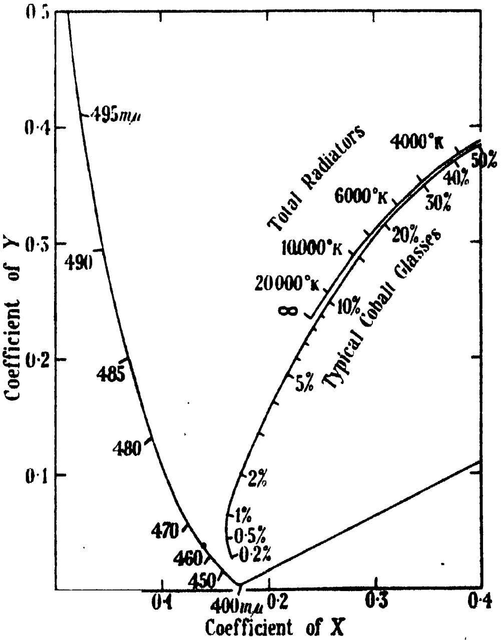


Figure 3. Colour-transmission relation for typical cobalt glass with 2848° K. light source.

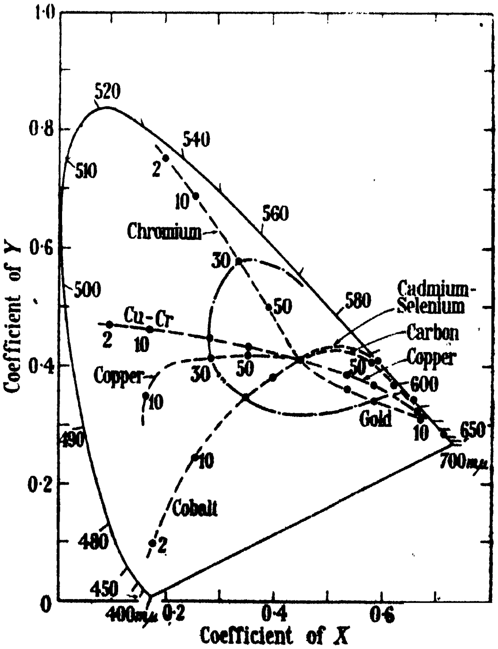


Figure 4. Colour-transmission relation for typical signalling glasses with 2848° K. light source.

will have lower transmission factors than uncoloured glasses, and a given colouring constituent will generally show a reproducible relation between the transmission factor and the colour coefficients. If the transmission factors for the six glasses in figure 1 are plotted against the coefficients of X and of Y in figure 2, this relation can be found by graphical interpolation, and a scale of transmissions can be drawn on the chromaticity diagram to show the colour-transmission locus of typical cobalt blue glasses with a 2848° K. light-source as in figure 3. Similar figures may be calculated from figure 1 for other light sources.

The colours given by other colouring constituents may be analysed in the same way, and figure 4 shows the colours given by the glasses commonly used in colour-light signals, with a light source at a colour temperature of 2848° K. (Holmes, 1937). The dashed lines are the loci of the colours given by varying thickness or concentration of the several colouring constituents, and the dots indicate the colours of glasses whose percentage transmission factor is written close by the dot. For example, a cobalt glass of 50% transmission factor with 2848° K. may be expected to transmit light whose colour is $0{\cdot}38X+0{\cdot}40Y+0{\cdot}22Z$. The chain-dot line in figure 4 connects all the points of 30% transmission, and it may be taken as a first approximation that no ordinary glass can give a colour outside this chain-dot line and also have a higher transmission factor than 30%, with a 2848° K. light source. This "maximum-transmission locus" may be compared with the maximum pigment colours calculated by MacAdam (1935), and it will be found that the yellow-orange and pure red glasses are not far removed from the theoretical maximum transmission for their colour, but the green and blue-green glasses give transmission factors much below the maximum for the same colour or, alternatively, give colours whose purity is much less than theoretically possible for the same transmission factor. It is of interest to see that the orange and red colours given by cadmium-selenium glasses can be matched by spectral colours. The explanation of this on the diagram is that these glasses absorb the short wave-length end of the spectrum completely, whilst transmitting the long wave-lengths with very little absorption, and if all wave-lengths less than about 540 mμ are absorbed, the transmitted light will be composed of wave-lengths whose colours are on the straight part of the spectrum locus and, therefore, the colour of the mixture will itself lie on the straight part of the locus and be matched by a spectral wave-length. A cadmium-selenium glass which absorbs below 540 mμ would have a transmission factor of about 55% and a colour of about $0{\cdot}585X+0{\cdot}414Y+0{\cdot}001Z$ with 2848° K., this colour being matched by the wave-length of 592 mμ in the yellow-orange part of the spectrum. The maximum theoretical transmission factor to give this colour is about 64%.

On the other hand, glasses which give green and blue colours will generally be of low saturation with artificial light. The light transmitted through a green glass will contain wave-lengths lying on the strongly curved part of the spectrum locus and, therefore, a highly saturated green colour can only be obtained by absorption of all except a narrow band of the spectrum. The light transmitted through a blue glass will usually be of relatively low saturation because of the relatively low energy level at the blue end of the spectrum from incandescent filament lamps, from which it is only possible to achieve a highly saturated blue colour by employing a glass of very low transmission factor as indicated in figure 3.

§4. THE MEASUREMENT AND CALCULATION OF TRANSMISSION FACTOR

The measurement and calculation of transmission factor is the first part of the colorimetry of glass, being simpler than the measurement of colour and yet bearing a close relationship to colour. Incidentally, the American word for transmission factor is "transmittance", and there is a risk of confusion with our word transmittance, which has a different meaning. We say that transmission factor is the ratio of the light leaving a glass to that incident upon it and that transmittance is the value which this ratio would have if there were no reflexion of light at the two air-glass surfaces. To a first approximation, the transmittance of flat glass is 1·08 times the transmission factor.

Lambert's law is strictly obeyed by all non-fluorescent glasses:—

$$I=I_0 \,.\, T=I_0 \,.\, 10^{-D},$$

where I is the intensity of the transmitted light, I_0 is the intensity of the incident light, T is the transmission factor, and D is the optical density.

The optical density D is the common logarithm of the reciprocal of the transmission factor T. The internal optical density d (sometimes written ID) bears the same relation to the transmittance t. Thus:

$$D=\log_{10}(1/T) \quad \text{or} \quad T=10^{-D},$$
$$ID=d=\log_{10}(1/t) \quad \text{or} \quad t=10^{-d}.$$

If r is the reflexion loss,

$$T=t \,.\, (1-r),$$
$$d=D+\log_{10}(1-r),$$
$$T=(1-r) \,.\, 10^{-d}.$$

Bouguer's law of variation of transmission factor with thickness (sometimes ascribed to Lambert) is obeyed provided the quality of the light is unchanged, as in truly neutral grey glasses or in monochromatic light:

$$I_x=I_0 \,.\, (1-r) \,.\, (t_1)^x=I_0 \,.\, (1-r) \,.\, 10^{-dx},$$

where I_x is the intensity after transmission through a thickness x and t_1 and d_1 are the transmittance and internal density for unit thickness.

Beer's law is not generally obeyed for variations in concentration of the colouring constituent, although a modified relation can be found as indicated below.

The thickness-conversion equations for transmission of monochromatic light through thicknesses x and y can be stated as follows:

$$t_x=(t_1)^x, \qquad (t_y)^x=(t_x)^y,$$
$$d_x=x \,.\, d_1, \qquad d_y=y \,.\, d_x/x.$$

Thus internal density is proportional to thickness for monochromatic light.

In terms of transmission factor,

$$T_y=(1-r) \,.\, [T_x/(1-r)]^{y/x}$$

or

$$\log T_y=\log(1-r)+[\log T_x-\log(1-r)] \,.\, y/x.$$

These relations are old established, but have only recently appeared in technical literature (Sharp, 1942, and McLeod, 1945). The last equation can be solved quickly by several methods, and the most accurate and quick method of converting from one thickness (x) to another thickness (y) is to use two slide rules, one set for the ratio y/x and the other set to read $\log_{10}[T/(1-r)]$. This second setting can be made by first calculating the reflexion loss from the Fresnel expression $(n-1)^2/(n+1)^2$, which leads to a value 0·92 for the factor $(1-r)$ if the refractive index n is 1·516, and then setting the linear scale, usually found on the back of the B- and C-scales of a slide rule, with its zero opposite to 92 on the D-scale. The linear scale will then be the internal density, which is proportional to thickness, and the D-scale will be the percentage transmission factor, T. A tabular form of the twin slide-rule method has been described by Gage (1937).

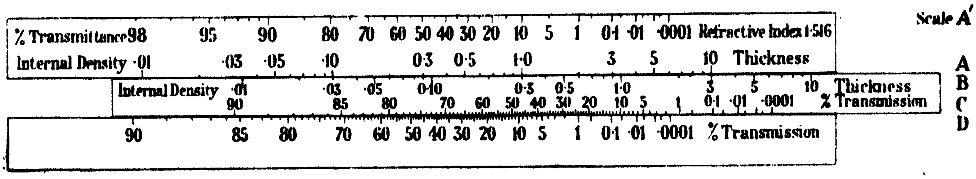

Scale A′
% Transmittance 98 95 90 80 70 60 50 40 30 20 10 5 1 0·1 ·01 ·0001 Refractive Index 1·516
Internal Density ·01 ·03 ·05 ·10 0·3 0·5 1·0 3 5 10 Thickness
A
Internal Density ·01 ·03 ·05 0·10 0·3 0·5 1·0 3 5 10 Thickness
B
90 85 80 70 60 50 40 30 20 10 5 1 0·1 ·01 ·0001 % Transmission
C
90 85 80 70 60 50 40 30 20 10 5 1 0·1 ·01 ·0001 % Transmission
D

Figure 5. Slide rule for transmission-thickness conversions.

It is not difficult to make a slide rule to solve the thickness-conversion equation directly for a given reflexion loss, and figure 5 is a sketch of such a rule which was made just before the war and which has been in almost daily use since that time. In figure 5, the slide is set for a 3 : 10 change in thickness as shown on the A- and B-scales, and the corresponding change in transmission factors may be read off the C- and D-scales, such as, for example, 50% transmission factor at 3 mm. thickness becomes 12% transmission factor at 10 mm. thickness. The A- and B-scales are actually a logarithmic ruling for the internal density, which is proportional to thickness, and the C- and D-scales are calculated from the equation. The A′-scale is the percentage transmittance corresponding to the internal density on the A-scale and the transmission factor on the D-scale. An ingenious circular form of this slide rule has recently been described by Vaughan (1944).

A very simple graphical method of calculating the effect of thickness changes is to use linear-log graph paper, suggested to the author by Dr. W. M. Hampton. In figure 6 the ordinates are the transmission factors on a logarithmic scale and the abscissae are thicknesses on a linear scale. The straight lines are the relations between thickness and transmission for two neutral glasses, both passing through 92% at zero thickness and elsewhere conforming strictly to the thickness-conversion equation above. This method has recently been published by Powell (1945). The curved line has been experimentally determined for a coloured

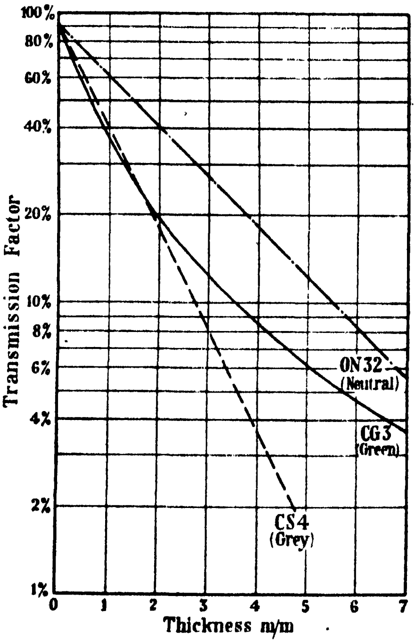


Figure 6. Graphical method for transmission-thickness conversions.

glass and it is found that all coloured glasses in which the quality or composition of the light is changing as it passes through greater thicknesses give concave curves of the type indicated. Some glasses, such as browns or blue-greens, give curves which are only slightly bent, and others, such as selenium rubies, give very strongly bent curves. The explanation of the curvature is that, as the thickness increases, the wave-lengths which are more strongly absorbed become of less importance and the light becomes relatively richer in the wave-lengths which are more freely transmitted, so that the transmittance per unit thickness is greater than it would be if the quality of the light had not changed, and so the slope of the line changes. Glasses of a given type will all show similar curves, and this method has proved most valuable in the routine control of coloured glass during the past fifteen years or so.

The departure from Beer's law may be represented in the same way, and the transmission factor for constant thickness but varying concentration may be shown as a curved line on a diagram similar to figure 6.

Emphasis has been placed on the arithmetical methods used in conversion from one thickness to another because these represent an important part of the glass-maker's colorimetric technique. The measurement of the transmission factor of glass presents no great difficulties, and it is the author's preference to employ visual methods exclusively. Photoelectric cells give quicker readings but tell lies without blushing. The reliability of flicker photometers is very dependent on the skill of the observer, and in extreme cases threefold errors have been obtained in measurement of the transmission factor of purple glasses with several observers. The most reliable method has been to use a photometer bench with a Lummer-Brodhun contrast head, two light sources of controllable colour temperature and a wide range of accurately calibrated glasses of all the common colours for use as comparison standards. The calibration is best done on a spectrophotometer, and this may of course be visual or photoelectric.

An interesting aspect of transmittance measurement is in the control of through-coloured stepped lenses (Fresnel lenses). There is a specification which defines a minimum transmittance (BSS 623–1940) and the method of measurement is to compare, on a photometer bench, the brightness of the photometer screen illuminated through the coloured glass lens and through a colourless glass plate with the brightness when the screen is illuminated through a colourless glass lens of the same pattern and a calibrated coloured glass plate. The light source is an opal lamp operated at the appropriate colour temperature and placed at the focus of the lens, conjugate to the photometer screen. Each of these assemblies is matched in turn against a coloured light on the other side of the photometer head, the properties of this light being unimportant except that there should be a reasonable colour match. The ratio of the two brightnesses is the same as the ratio of the unknown transmittance of the coloured lens to the known transmittance of the coloured glass plate, and thus a measurement which might be very complex becomes a matter of simple routine. For some types of glass the limiting colour stated in the specification can be correlated with a maximum transmittance and, as it is often found that there is a close correlation between weight or thickness and transmittance in a single batch of glasses, it is not unusual to determine the limits by careful colorimetric technique and then to carry out the routine examination of each glass with a spring balance or dial gauge. Border-line glasses would of course be subjected to colorimetric examination.

§ 5. THE MEASUREMENT AND CALCULATION OF COLOUR

The transparent nature of glass emphasizes that, in common with other coloured materials, it possesses no colour of its own, but only shows colour by its selective absorption of light passing through it. Unlike a surface colour, glass is not usually looked at, but is looked through, and it is immediately obvious that we have to measure the colour of the combination of the glass and the light source. This point is particularly important in the colorimetry of signal glasses which are used with light sources of widely varying colour-temperature and misconceptions are liable to arise.

The simplest method of checking that the colour of a glass is between defined limits is to compare the test glass with limit glasses or with a calibrated standard. The photometer bench with lamps of adjustable colour temperature and calibrated standard glasses of the same type as the test glass will give sufficient accuracy for most purposes. Where interpolation is required to a higher accuracy than by visual estimation between two colours, it is common practice to use a wedge of glass which is calibrated for colour along its length and to match the test glass against the appropriate thickness of the wedge.

Any type of colorimeter can be used for measuring the colours of glasses if there is some attachment enabling it to measure the colour of light. The Hilger-Guild instrument or the Donaldson instrument are both excellent for this purpose and the Lovibond Tintometer is very good except for the purest colours, which may require the addition of a neutral shade to bring them within the range of the instrument. The type of "colorimeter" which ought to be called an "absorptiometer" is not, of course, suitable for direct measurement of colour, but it may be used to give reliable results by the method of abridged spectrophotometry if it is suitably calibrated, preferably by reference to a known glass of the same type as the test glass.

The best method of colour measurement is to determine the transmission factor throughout the visible spectrum, because this gives the whole relevant information about the glass, and this information can be readily handled by calculation. The author's usual method is to calculate the trichromatic coefficients and the transmission factor from the spectrophotometric data for a series of thicknesses, leading to the colour-transmission relation for the glass as shown in figure 3 for cobalt glass. The only practicable experimental method for obtaining data such as this is the spectrophotometric method.

From a knowledge of the typical spectrophotometric transmission curve, the effect of change of the colour temperature of the source can be readily calculated, and curves such as those in figure 4 can be drawn for other light sources. This is particularly valuable for glasses for coloured light signals which may have any illuminant from an oil flame at 1900° K. to an arc lamp at 3500° K. and in which the effect of change of the illuminant may be greater than the whole tolerance allowed for the colour of the glass itself (Holmes, 1937). Figure 7 shows the colours and transmission factors of Aviation Green glass (Chance CG6) for three light sources, calculated from a single spectrophotometric curve by the methods described below.

In the calculation of colour it is general practice to employ the CIE trichromatic system, although a number of users of coloured glass continue to describe its colour in terms of wave-length and saturation. The author's calculations are on the weighted ordinate method rather than the selected ordinate method (Hardy, 1936) because the former gives a higher accuracy with less labour. The weighted ordinate method was described by Smith and Guild in 1931 and consists of two stages for each of the three primaries, the first stage being to multiply the transmission factor (T_λ) by the energy level of the light source (E_λ) and by the distribution coefficient ($\bar{x}_\lambda$, etc.) recommended by the CIE in 1931, making this calculation for each wave-length at intervals of, say, 0·01 micron through the spectrum, and the second stage being to add the products to give the trichromatic equation for the colour. The first stage can be greatly simplified by constructing a permanent table of products, either by calculating machine or by careful use of a slide rule, and the second stage can then be done by adding selected products on an adding machine, so avoiding the necessity for laborious multiplication. Table 1 shows some of the entries in the master table for X, Y and Z for Illuminant A. In the master table, the required product ($T \,.\, E \,.\, \bar{x}$ etc.) has been calculated for each percentage transmission from 1% to 90% and for each wave-length at intervals of 0·01 micron, based on the figures given by Smith (1934). The appropriate products are chosen from each column and added by machine to give the coefficients of the trichromatic equation. Although the products are only worked out for each 1% transmission, it is simple to make calculations for decimal percentages by moving the decimal point.

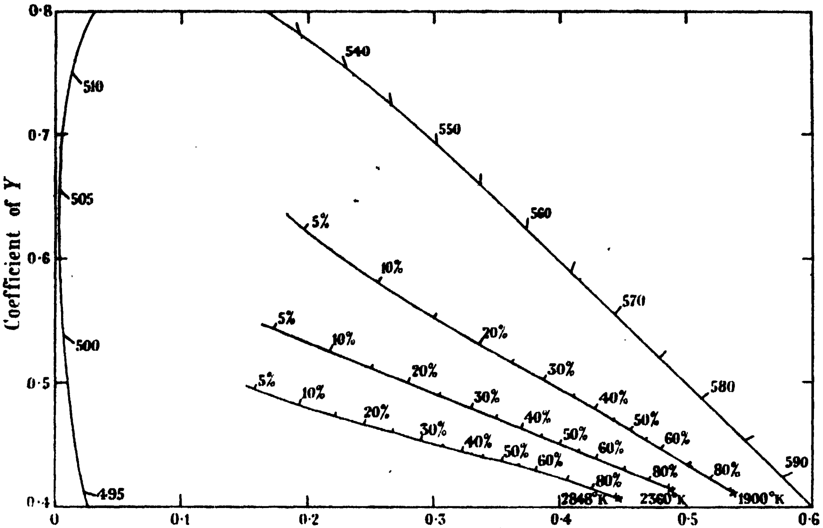
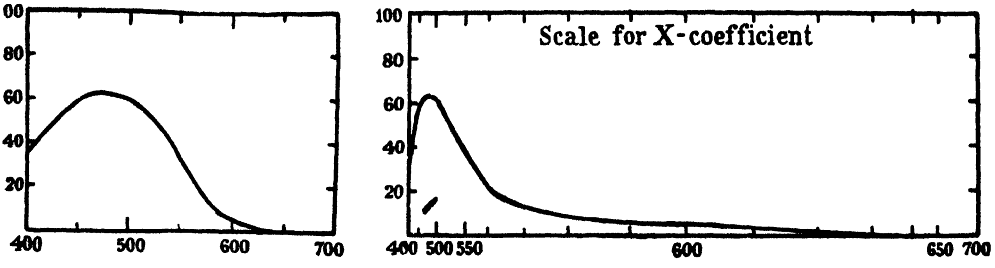


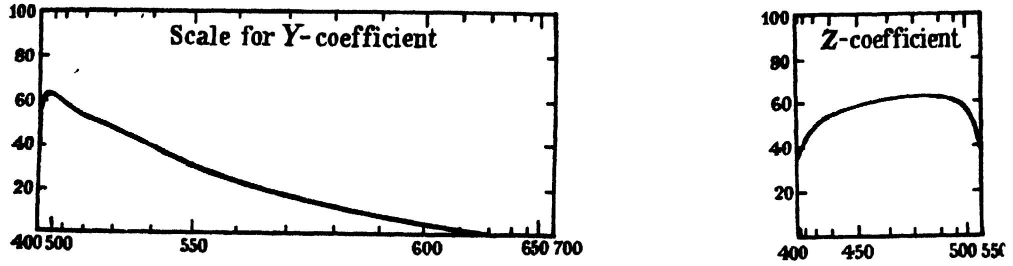
Figure 7. Colour-transmission relations for aviation green glass (Chance CG6).

The weighted ordinate method can also be applied graphically, using scales which have been published in an earlier paper (Holmes, 1935) as illustrated in figure 8. Figure 8 (*a*) shows the curve of transmission factor and wave-length for a blue-green glass, plotted with linear scales in the ordinary way. Figures 8 (*b*), 8 (*c*) and 8 (*d*) show the same curve plotted with a linear scale of transmission factors but with non-linear scales of wave-lengths. The spacing of each wave-length scale has been calculated according to the energy distribution of the light source (in this case, Illuminant A or 2848° K.) and the distribution coefficients of the standard observer for each of the three primaries. It follows that the area under each of the three curves is proportional to each of the three trichromatic coefficients of the colour and if the areas are reduced to unit sum, the result will be the unit trichromatic equation. The area under the curve for the Y-coefficient is, of course, the transmission factor of the glass. The advantages of the method are firstly that you can see what you are doing and secondly that a reasonably accurate result can be obtained with very few observed points on the transmission curve. The method is quick if permanent ruled charts are kept, but it is subject to all the usual errors of the measurement of area by a planimeter. Incidentally, it may be argued that this is a graphical form of the selected ordinate method, rather than the weighted ordinate method, and it may be derived from either.



(*a*) Graphical calculation of CIE coefficients for Illuminant A.
Source: $0{\cdot}448X + 0{\cdot}407Y + 0{\cdot}145Z$.
Filter: Chance OB2 at 2 mm.

(*b*) Total area 109·85. Area under curve 12·4.



(*c*) Total area 100·00. Area under curve 22·0.

(*d*) Total area 35·58. Area under curve 20·6.

Figure 8. Graphical method for colour calculations.
Unit equation $0{\cdot}225X + 0{\cdot}400Y + 0{\cdot}375Z$. Transmission factor 22·0%.

The idea of a unit equation is sometimes rather difficult at first for a student and it may be suggested that a "percentage equation" is easier to understand. In the example quoted in figure 8 the light transmitted by the glass is represented by the three areas: $x' = 12{\cdot}4$, $y' = 22{\cdot}0$, $z' = 20{\cdot}6$.

If the total is regarded as 100%, which is immediately understandable to any student, it is clear that this corresponds to 22·5% of X, 40% of Y and 37·5% of Z, and he may then see that this is another way of expressing the unit equation:

$$C = 0{\cdot}225X + 0{\cdot}400Y + 0{\cdot}375Z.$$

The unit equation therefore represents the "proportions" of a colour rather than its "amount "or relative luminosity.

The relation between unit equations of a colour with different primaries may also cause difficulty to a student, and a diagram such as figure 9 may help in the understanding of the transformation equations. This diagram shows the unit triangle of the *RGB* primaries of the author's colorimeter (Holmes, 1935) superimposed on the usual rectangular *XYZ* diagram in such a way that the unit *RGB* equation and the unit *XYZ* equation of any colour both plot at the same point in the two scales. For example, the unit equations for Illuminant B plot at:—

$S_B = 1/3 \,.\, R + 1/3 \,.\, G + 1/3 \,.\, B$ in the *RGB* triangle, and
$S_B = 0{\cdot}348X + 0{\cdot}353Y + 0{\cdot}300Z$ in the *XYZ* triangle.

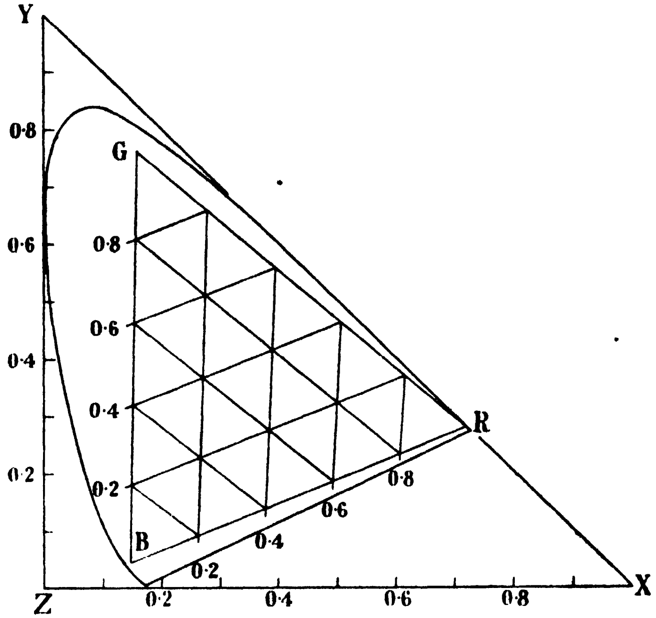


Figure 9. Relation between unit *RGB* triangle and unit *XYZ* triangle.

The *RGB* triangle has straight sides and it may be sub-divided by straight lines, although the scales are not quite evenly spaced. If monochromatic stimuli were employed (Smith and Guild, 1931, page 78) the *RGB* triangle would be nearly equilateral, but would have less evenly spaced scales for the coefficients of the three primaries.

The transformation equations to be employed in producing a chart such as figure 9 may be derived as follows:—

The calibration of the colorimeter gives three equations for the instrumental primaries:

$$R = x_1X + y_1Y + z_1Z,$$
$$G = x_2X + y_2Y + z_2Z,$$
$$B = x_3X + y_3Y + z_3Z.$$

In these equations, x_1 etc. are the coefficients determined by the calibration and usually the sum of all nine coefficients is 3·000.

A colour C can be represented by unit equations:

$C = rR + gG + bB$ on the instrumental (RGB) system.

$C = xX + yY + zZ$ on the C.I.E. (XYZ) system.

These equations may be reduced to a form which is easy to handle arithmetically:

$$x = \frac{r(x_1 - x_3) + g(x_2 - x_3) + x_3}{r[(x_1 + y_1 + z_1) - (x_3 + y_3 + z_3)] + g[(x_2 + y_2 + z_2) - (x_3 + y_3 + z_3)] + (x_3 + y_3 + z_3)},$$

$$y = \frac{r(y_1 - y_3) + g(y_2 - y_3) + y_3}{r[(x_1 + y_1 + z_1) - (x_3 + y_3 + z_3)] + g[(x_2 + y_2 + z_2) - (x_3 + y_3 + z_3)] + (x_3 + y_3 + z_3)}.$$

For the author's colorimeter, the transformation equations are:

$$x = \frac{0{\cdot}594r + 0{\cdot}001g + 0{\cdot}150}{0{\cdot}011r - 0{\cdot}074g + 1{\cdot}021},$$

$$y = \frac{0{\cdot}242r + 0{\cdot}675g + 0{\cdot}046}{0{\cdot}011r - 0{\cdot}074g + 1{\cdot}021}.$$

These equations are easily solved on a slide rule.

In the testing of signal glasses, it is sometimes desired to compare the result of a colorimeter measurement with the limits stated in a specification in terms of areas on the XYZ diagram. A chart such as figure 9, or a portion of it, drawn on an enlarged scale, enables the experimental results to be plotted on the instrumental (RGB) system and the comparison with the specification on the XYZ system can then be seen, no intermediate calculation being necessary. Alternatively, the reverse transformation may be calculated and the specification limits converted to the instrumental system and then the comparison may be made on a simple RGB diagram. The choice between the two methods depends on the frequency of making comparisons with any particular specification, the latter method being preferable for routine work.

§ 6. THE DESIGN OF COLOUR FILTERS

The requirements for a glass colour filter may be stated in terms either of the colour which is required with a given illuminant or of the spectrophotometric transmission curve. The former statement of the problem is usually easier to satisfy, because two glasses can often be found on opposite sides of the required colour and a match can be obtained by subtractive mixing. As an example of this, the Aviation Green glass in figure 7 was obtained by a mixture of copper oxide and chromium oxide, data for each of which are shown in figure 4. In obtaining the match, an hour or so of calculation by trichromatic methods may save days of experimental meltings. Allowance must be made for the base glass employed, that is to say, for the network and modifiers into which the colouring ions are to be admitted, for the conditions of oxidation or reduction in the melting process for traces of impurities or of oxides which react with the colouring oxides and then the degree of difficulty depends on how closely the required colour has to be matched.

If a transmission curve has to be matched, the problem is more complex. The transmission curves of most of the simple colouring oxides on the simple base glasses are known and some laborious arithmetic may enable a fairly close match to be obtained by subtractive mixing of two or three transmission curves. The arithmetic is greatly simplified if the curves are in terms of internal density, when the process of subtractive mixing is one of simple arithmetical addition at each wave-length, rather than multiplication. If the calculation is done graphically, the use of proportional dividers for scaling from a curve of internal density (or of transmission on a logarithmic scale) and wave-length enables any given thickness or concentration of each colouring oxide to be used in the calculation. Having obtained an approximate answer, you try it, and by successive approximations, gradually approach your target. There is scope for considerable personal skill in choosing the right starting point and estimating the corrections to be applied.

As an example of what can be done, figure 10 shows the transmission of some colour-temperature conversion filters; the curved dashed line represents the type of glass filter generally available in this country before the war, and the straighter line represents a filter which has been developed during the war. This diagram is plotted on an unusual scale suggested by Gage (1933) with the transmission on a logarithmic scale and the wave-length on a reciprocal scale, chosen because a perfect colour-temperature conversion filter, based on the Wien equation, would plot as a straight line.

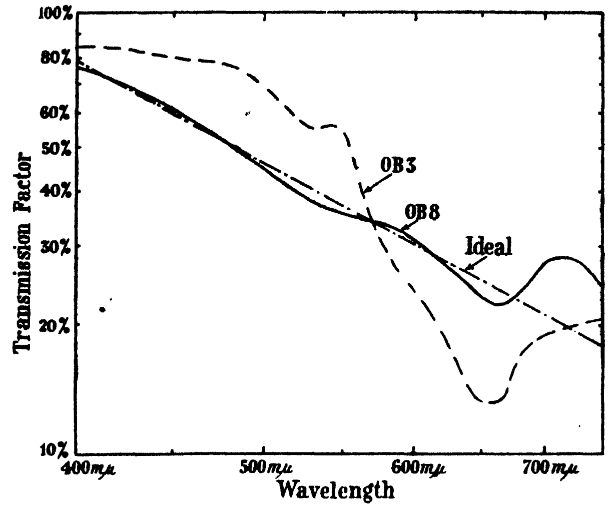


Figure 10. Transmission curves for colour-temperature conversion filters.

§ 7. THE COLOURS OF BLOOMED LENSES AND OF POLARIZED LIGHT

It is usual to estimate the reflexion factor of a bloomed surface, as used on a lens in an optical instrument, by looking at its colour; and the relation between these two provides a rather interesting use of colorimetry. If a film of low index is put on the surface of a glass, the combined reflexion factor of the two surfaces of the film is less than that of the single surface of the glass and, if the film is thin enough for destructive interference to take place, the reflexion factor can be reduced very considerably at some part of the spectrum. The reflected light is therefore coloured and the colour and the total reflexion factor are related to each other. Figure 11 shows the curves of reflexion factor calculated by classical theory for different values of the optical thickness (nt) of a film of refractive index (n) 1·30 on a base glass of refractive index 1·69. The reflexion factor of one surface of the base alone is 6·8% and the addition of the film would reduce this to 3·4% if no interference took place. If the interference is destructive, as for a wave-length of 0·54μ and an optical thickness of 0·135μ, the reflexion factor is zero, but if the interference is reinforcing, as for a wave-length of 0·50μ and an optical thickness of 0·25μ, the reflexion factor is 6·8%. The curves in figure 11 show that thin films ($nt = 0{\cdot}05\,\mu$ or $0{\cdot}10\,\mu$) are yellow-orange-red by reflected light, that a film of about $0{\cdot}135\,\mu$ optical thickness will have a low total reflexion factor and a purple colour and that thicker films ($nt = 0{\cdot}2\,\mu$ or $0{\cdot}25\mu$) are blue-green by reflected light.

If the trichromatic coefficients and total reflexion factors are calculated for a number of thicknesses and plotted as in figure 12 it will be seen that the minimum total reflexion factor occurs at about $0{\cdot}14\,\mu$ for the optical thickness (nt) of the film and that there is a very rapid change in the trichromatic coefficients near this thickness. It follows that there can be a wide variation in colour without any significant variation in reflexion factor. This is also shown in figure 13, where the solid line shows the colours and total reflexion factors calculated from curves as in figure 11 and the dashed line shows the colours of less saturation which are obtained if the interference is not completely destructive as, for example, when the refractive index of the film is not the square root of that of the base glass. Considering the solid line, a reflexion factor of 0·3% may be obtained with a red-coloured (thin) film or a blue-coloured (thick) film and it follows that films of intermediate colours will have lower reflexion factors. The same general conclusions may be applied to the dashed line, which shows the colours likely to be obtained from real films in which the coloration produced by destructive interference is diluted by white light.

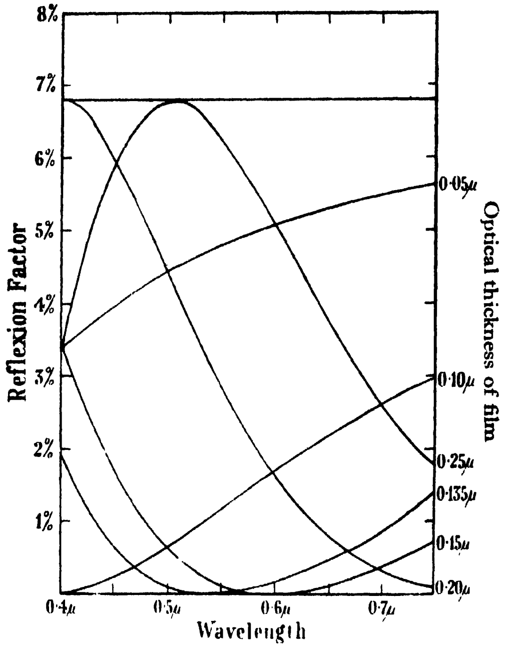


Figure 11. Calculated reflexion factor curves for bloomed glass. (Several film thicknesses.)

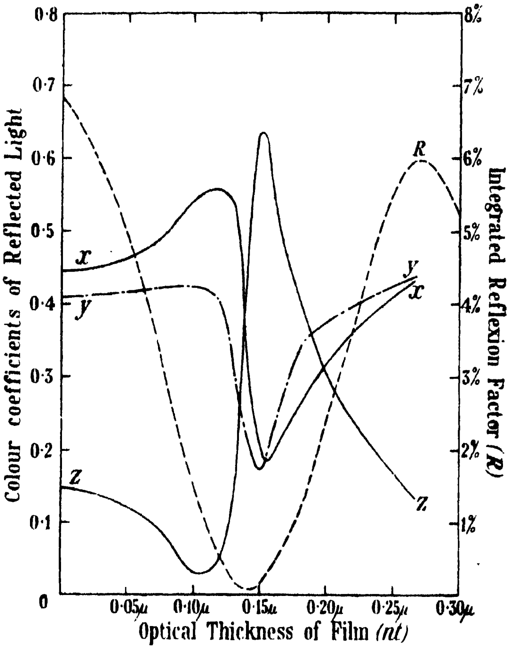


Figure 12. Calculated relation between trichromatic coefficients and film thickness.

Figure 13 is very similar to the diagram in Dr. Wright's *The Measurement of Colour* (1945), based on the work of Baud and Wright (1930) in which photoelastic colours are illustrated. These colours are obtained by an entirely different physical process but it so happens that both interference and polarization give a cosine law for the intensity distribution and the chromaticity diagrams are therefore similar.

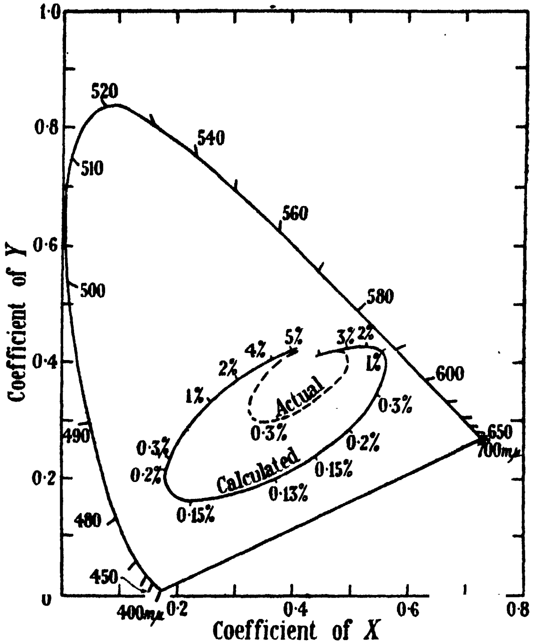


Figure 13. Colours and reflexion factors for bloomed glass with 2848° K. source.

This leads to another use of colour in the manufacture of glass, namely the inspection of glass for internal strain by polarized light.

If the retardation is of the order of a wave-length, there is a very rapid change in colour for a relatively small change in retardation or, because the retardation in a strained glass is proportional to the stress, for a relatively small change in stress. This is similar to the rapid change in colour for a small change in film thickness on bloomed lenses. In a strain-viewer or polariscope for examining the annealing of glass, a tint plate is used, which has one wave-length retardation for yellow-green light (0·55 μ), and a small change in this retardation due to a small strain in the glass will change the colour of the transmitted light and, as it appears that the changed colour lies actually in the strained glass, the regions of maximum strain are readily found and the amount of this strain may be estimated.

Opinions differ as to the best colour to use for the tint plate—whether it should be 0·54 μ or 0·57 μ retardation—but actually this is a simple problem in colorimetry and capable of an exact solution. The diagram given by Dr. Wright shows the colours given by different retardations with a light source at about 3000° K. and the scale of retardation is most widely spaced at about 0·56 μ. If daylight were used, the effective wave-length of the white light would be less and 0·55 μ or 0·54 μ might be better. The exact optimum value can be obtained by calculating the colours with the particular light source involved and plotting them on a uniform chromaticity scale (Holmes, 1940) as in figure 14. The optimum retardation is at the point on the curve where a change in retardation causes the maximum linear displacement of the colour as plotted, and the answer from Baud and Wright's work is 0·572 μ. This is one of the uses to which a uniform chromaticity scale may be put with complete confidence because small departures from true uniformity will not affect the result by an appreciable amount.

§ 8. THE SPECIFICATION OF COLOURED GLASSES

The specification of coloured glasses has changed very considerably during the last twenty years, and the story of glasses for coloured light signals can be used as an example of the developments which have taken place. Over twenty-five years ago it was usual to have a target colour and then to have a certain amount of argument as to whether the glass supplied was the same as the target. The various railways used different target colours—the Scottish Railways used emerald green, the Welsh used purple or violet, the Great Western used blue-green (with limit glasses) and other lines used other varieties of green. Our knowledge of the colorimetric properties of the glasses was no more than the end-point of a spectrum photograph or, in a few cases, the total transmission factor. When the railways merged in 1922, the four main lines agreed to adopt the red and green glasses proposed by a Board of Trade Commission for ships' lights, and the light and dark limits were specified in terms of spectrophotometric transmisssion curves.

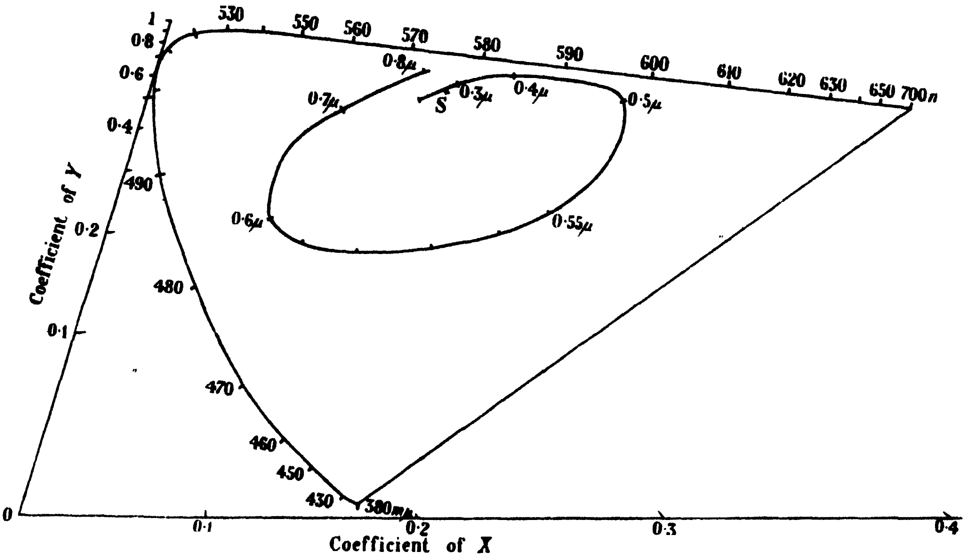


Figure. 14. Colours and retardation for polarization colours (after Baud and Wright).

The glasses shown were well suited to oil flame illuminants and limit glasses were specially made and precisely specified. In 1928 an orange range was introduced, the choice of the colour being based on laboratory experiments at the N.P.L. (Guild, 1928) on the risk of confusion with red or with white signals. Limit glasses were chosen and recorded in terms of their spectrophotometric transmission curves.

In 1933 the British Standards Institution set up a representative committee of Signal Engineers of Railways, representatives of the Ministry of Transport, Signal Makers, National Physical Laboratory, and others, to prepare a specification for coloured railway signal glasses and full use was made of the trichromatic methods of description of colour which had been agreed by the C.I.E. in 1931. A series of full-scale tests showed which glasses were too dark or too light for each colour and for each type of signal and the limiting glasses were measured with a standard light source (2360° K.) at the N.P.L. The colours were plotted on a chromaticity chart and it was possible to prescribe the areas on the chart within which a glass must lie if it was to be acceptable in service. The specification

BSS 623 was issued in 1935 and was revised in 1940, when only slight changes were necessary.

In 1937 laboratory-scale experiments on the reliability of recognizing any particular colour as red or green had been commenced by the author; and the 1940 revision of the specification was based on the results available at that time as well as on full-scale experience. These colour-recognition experiments have since been completed (Holmes, 1941) and a new specification for coloured glasses for all types of signal—railway, marine, avaition, street traffic—is now being prepared,

Table 1. Abstract from table of products for colour calculation with Illuminant A (2848° K.)

Wavelength (mμ)	400	450	500	550	600	650	700	750
Transmission factor	Products $T.E.\bar{x}$ to give coefficient of X							
100%	1·93	103·20	2·69	373·29	1271·03	434·47	20·67	0·63
90	1·74	92·88	2·42	335·96	1143·92	391·02	18·60	0·57
80	1·54	82·56	2·15	298·63	1016·82	347·58	16·54	0·50
70	1·35	72·24	1·88	261·30	889·72	304·13	14·47	0·44
60	1·16	61·92	1·61	223·97	762·62	260·68	12·40	0·38
50	0·97	51·60	1·35	186·65	635·52	217·24	10·34	0·32
40	0·77	41·28	1·08	149·32	508·41	173·79	8·27	0·25
30	0·58	30·96	0·81	111·99	381·31	130·34	6·20	0·19
20	0·39	20·64	0·54	74·66	254·21	86·89	4·13	0·13
10	0·19	10·32	0·27	37·33	127·10	43·45	2·07	0·06

Wavelength (mμ)	400	450	500	550	600	650	700	750
Transmission factor	Products $T.E.\bar{y}$ to give coefficient of Y (Divide sum by 20 to give total transmission for 2848° K.)							
100%	0·05	11·67	179·56	857·07	754·60	163·89	7·44	0·21
90	0·04	10·50	161·60	771·36	679·14	147·50	6·70	0·19
80	0·04	9·34	143·65	685·66	603·68	131·11	5·95	0·17
70	0·04	8·17	125·69	599·95	528·22	114·72	5·21	0·15
60	0·03	7·00	107·74	514·24	452·76	98·33	4·46	0·13
50	0·02	5·84	89·78	428·54	377·30	81·94	3·72	0·10
40	0·02	4·67	71·82	342·83	301·84	65·56	2·98	0·08
30	0·02	3·50	53·87	257·12	226·38	49·17	2·23	0·06
20	0·01	2·33	35·91	171·41	150·92	32·78	1·49	0·04
10	0·00	1·17	17·96	85·71	75·46	16·39	0·74	0·02

Wavelength (mμ)	400	450	500	550	600	650	700	750
Transmission factor	Products $T.E.\bar{z}$ to give coefficient of Z							
100%	9·16	543·91	151·32	7·49	0·96	0·00	0·00	0·00
90	8·24	489·52	136·19	6·74	0·86	0·00	0·00	0·00
80	7·33	435·13	121·06	5·99	0·77	0·00	0·00	0·00
70	6·42	380·74	105·92	5·24	0·67	0·00	0·00	0·00
60	5·50	326·35	90·79	4·49	0·58	0·00	0·00	0·00
50	4·58	271·95	75·66	3·75	0·48	0·00	0·00	0·00
40	3·66	217·56	60·53	3·00	0·38	0·00	0·00	0·00
30	2·75	163·17	45·40	2·25	0·29	0·00	0·00	0·00
20	1·82	108·78	30·26	1·50	0·19	0·00	0·00	0·00
10	0·92	54·39	15·13	0·75	0·10	0·00	0·00	0·00

using this work and some parallel experiments on aviation signals made in 1939 at the Royal Aircraft Establishment (Hill, 1939). The method of preparing the specification is to start with the data from practical experience, record it in trichromatic terms and to use the results of laboratory experiments to rule out the abnormal or irrelevant data, to interpolate or to extrapolate and to make deductions from the practical data. The conclusions can be expressed in terms of the properties of the satisfactory glasses when measured with a standard light source (Illuminant A—2848° K.) by a standard method and recorded on a chromaticity diagram. The correlation of the data presents a most intriguing problem in the interpretation of colour recognition as well as in the calculation of colour and of the change of colour with changed conditions; and the work before this B.S.I. committee is one of the most promising possibilities of advance of colorimetric technique in the glass industry.

REFERENCES

BAUD, R. V. and WRIGHT, W. D., 1930. *J. Opt. Soc. Amer.*, **20,** 381.

British Standard Specification, 623–1940. "Colours for Signal Glasses for Railway Purposes."

GAGE, H. P., 1933. *J. Opt. Soc. Amer.*, **23,** 46; 1937. *Ibid.*, **27,** 160.

GUILD, J., 1928. *Proc. Int. Conf. Illum.*, p. 862.

HARDY, A. C., 1936. *Handbook of Colorimetry* (Mass. Inst. Tech., Cambridge, Mass., U.S.A.).

HILL, N. E. G., 1938. *Report No. E & I* 1159.

HOLMES, J. G., 1935. *Proc. Phys. Soc.*, **47,** 400 ; 1937. *Proc. Inst. Lighthouse Conf. Berlin*, pp. 99–101 (German text) and pp. 77–79 (English text) ; 1940. *Proc. Phys. Soc.*, **52,** 359 ; 1941. *Trans. Illum. Engng. Soc., London*, **6,** 71.

MACADAM, D. L., 1935. *J. Opt. Soc. Amer.*, **25,** 361.

MCLEOD, J. H., 1945. *J. Opt. Soc. Amer.*, **35,** 185.

MOREY, G. W., 1938. *The Properties of Glass* (New York : Reinhold Publishing Corporation), p. 34.

POWELL, H. E. 1945. *J. Opt. Soc. Amer.*, **35,** 428.

SCHOLES, S. R., 1945. *Glass Ind.*, **26,** 417.

SHARP, D., 1942. *Glass Ind.*, **23,** 331.

SMITH, T., 1934. *Proc. Phys. Soc.*, **46,** 372.

SMITH, T. and GUILD, J., 1931. *Trans. Opt. Soc.*, **33,** 73.

VAUGHAN, T. C., 1944. *Glass Ind.*, **25,** 259.

WEYL, W. A., 1944. *J. Soc. Glass Tech.*, **28,** 158.

WRIGHT, W. D., 1945. *The Measurement of Colour* (London : Adam Hilger Ltd.), p. 196.

ULTRA-VIOLET BANDS OF Na_2

BY S. P. SINHA,

Imperial College of Science and Technology

MS. received 4 *December* 1946

ABSTRACT. The ultra-violet bands of Na_2 have been studied in absorption and emission. In absorption, by varying the conditions of temperature and pressure the bands have been photographed from λ3640 A. to λ2500 A. They are considered to belong to seven different systems, of which, however, only three are well developed. The classification of the other four systems is tentative. It has been shown that the Na_2 ultra-violet bands measured by Walter and Barratt (1928), which were analysed into five different systems by Weizel and Kulp (1930), may really be considered to belong to two systems only, corresponding to systems 1 and 3 of the present investigation, which are the most intensely developed ones.

The emission bands are weakly developed and have been observed only in the regions λλ3370–3180 A. and λλ3070–2960 A. These correspond to the strong bands of systems 1 and 3 in absorption.

§ 1. INTRODUCTION

THE diatomic molecule of sodium is known to possess two systems of bands in the visible region, one of which lies in the yellow-red and the other in the green. Extensive studies of these two systems have been made by several authors: the vibrational structure of the yellow-red bands by Fredrickson and Watson (1927) and Fredrickson and Stannard (1933), their rotational structure by Fredickson (1929); the vibrational structure of the green bands by Loomis and Nusbaum (1932), and their rotational structure by Loomis and Wood (1928). The vibrational and rotational constants of the molecule for the states involved in these two systems of bands are thus known to a high degree of accuracy.

The sodium molecule is known also to possess some bands in the ultra-violet region in absorption. These were first observed by Wood (1909). Walter and Barratt (1928) also obtained these bands, and their measurements were arranged into five different systems by Weizel and Kulp (1930). Further investigations in this region were carried out by Kimura and Uchida (1932) who, using the light from the crater of a carbon arc, photographed the absorption spectrum of sodium on a Hilger E quartz spectrograph and arranged the bands thus obtained into as many as six systems. Although Kimura and Uchida attributed a larger number of bands to each system than did Weizel and Kulp, and since they also obtained a large number of bands at shorter wave-lengths not observed by Walter and Barratt, some of their systems appeared to be rather incomplete, and it seemed worth while to photograph the bands and attempt a new analysis. Accordingly further observations have been made both on the absorption and emission spectra. The results of measurements and conclusions are described in the following sections.

§ 2. SOURCE, APPEARANCE OF THE SPECTRUM AND MEASUREMENTS

(a) *Bands in absorption*

The details of the apparatus used in the absorption experiment during the present investigation have been described elsewhere (Bhattacharya and Sinha, 1943). Sodium, freed from the oil in which it was stored, was put in a steel cell kept inside an electrically heated steel tube provided with water-cooled quartz windows, and light from a hydrogen discharge tube was used as source for the continuous radiation. The hydrogen gave a perfect continuum except for some OH bands near λ3100 A., which were eliminated during measurements. The pressure inside the absorption tube could be varied by introducing nitrogen gas from a cylinder. The presence of nitrogen inside the absorption chamber has been found to facilitate the appearance of Na_2 bands in the ultra-violet region.

Preliminary investigations were made with a Hilger Intermediate quartz spectrograph, and the optimum conditions of temperature and pressure for the bands to develop satisfactorily in different regions were noted. Using these values of temperature and pressure, the spectrum was next photographed in an E_1 quartz instrument.

At about 705° c., when the total pressure in the chamber is about 5 cm. of mercury, the bands are strongest in the region from λ 3280 A. to λ 3450 A., and the λ 3303 A. line of sodium is not very broad. On increasing the temperature and pressure, nearly the whole of this region is continuously absorbed and bands appear at wave-lengths greater than λ 3450 A. Bands also appear quite strongly between λ 2880 A. and λ 3100 A., and fainter bands extend up to about λ 2500 A. The conditions that favour the satisfactory development of bands in different regions are:

Temperature (°c.)	750	850	900
Total pressure in the tube (cm. Hg)	5	15	25
Region	λλ 3280–3450 A.	λ 3400–3640 A. λ 2880–3280 A.	λ 2500–2880 A.

and measurements included in table 1 in different regions correspond to these values of temperatures and pressures.

Measurements were made of all the band-heads obtained on both the spectrographs. Those measured on the smaller dispersion instruments are, however, not presented here except in the regions λ 3639 A. to λ 3542 A. and λ 2837 A. to λ 2496 A., in which regions bands on the spectrograms obtained on E_1 were too faint to be measured.

Bands were nearly all degraded to the red, although in some cases the heads were not very prominent under the microscope. In some cases it was difficult to decide which way the band was degraded, and in such cases (marked with asterisks) the centres of absorption were measured. Comparison of readings made with different spectrograms on the E_1 instrument showed that the measurements could be relied upon up to ± ·3 A. in most cases, although in some they were uncertain up to as high as ± ·5 A. Most of the bands measured could be classified and are given in table 1. The few remaining ones were mostly faint. Intensities given in table 1 are visual estimates on a scale of 10, and have been taken from different spectrograms for the different regions separated by horizontal lines in table 1.

(b) *Bands in emission*

The main purpose of studying the ultra-violet bands of sodium in emission was to settle the problem of arranging them into different systems.

The source was a discharge tube of simple design and essentially not different from that used by Wood and Galt (1911) and Kimura and Uchida (1932) for studying the visible and the ultra-violet bands of Na_2 respectively. The tube was of Pyrex glass 50 cm. long and 3·5 cm. in diameter and having three side tubes, two to carry the electrodes and the third to be connected to the vacuum pump. One end of the tube was closed and, after introducing some sodium freed from oil in the centre of the tube, a quartz window was sealed to the other end. The central portion of the tube over a length of about 12 cm. could be heated electrically from outside up to about 500° c. No cooling near the window was necessary.

The tube was first excited by an induction coil and the spectrum was photographed for different potentials over a range corresponding to about 4 to 10 inches of spark gap in air. The temperature was between 300° and 500° C. Although the visible bands appeared, there was no trace of any band in the ultra-violet. The tube was then excited by a transformer capable of giving a large current density though not a large potential. Under this excitation, at about 500° C., when the pressure in the tube was about 1 mm. of mercury, a brilliant yellow light appeared. The visible bands could be photographed quite readily, and in less than an hour's exposure some ultra-violet bands also appeared on the plate with the medium quartz instrument.

The measurements are given in table 2. The bands appear in three distinct regions: (1) λ 3370 A. to λ 3180 A., (2) λ 3070 A. to λ 3000 A. and (3) λ 2960 A. to λ 2900 A. In appearance they resemble the absorption bands on the Intermediate quartz spectrograph, except that the latter are much more intense.

Table 1. Ultra-violet absorption bands of Na_2

λ_{air} (A.)	Intensity	v',v''	System	λ_{air} (A.)	Intensity	v',v''	System
3639	1	2,16	1	3340·6	10	3,0	1
3619·5	1	2,15	1	3337·4	2	6,2	1
3600	1	1.13	1	3333·9	4	5,1	1
3580	2	1,12	1	3327·9	10	4,0	1
3561·5	2	0,10	1	3325·2	1	3,15	2
3542	2	0,9	1	3316·0	8	5,0	1
3510·2	2	1,8	1	3312·7	2	4,15	2
3505·4	2	0,7	1	3309·1	4	7,1	1
3479·6	2	2,7	1	3304·1	3	6,0	1
3474·6	3	1,6	1	3298·9	4	8,1	1
3452·9	4	4,7	1			4,14	2
		0,4	1	3295·8	2	3,13	2
3448·7	2	3,6	1	3291·8	5	7,0	1
3443·5	2	2,5	1	3287·3	1	5,14	2
3439·0	5	1,4	1	3283·0	1	4,13	2
3434·8	6	4,6	1	3280·4	4	8,0	1
3424·8	2	2,4	1			3,12	2
3416·4	4	0,2	1	3278·5	1	7,15	2
3412·5	2	3,4	1	3274·6	4	10,1	1
3409·7	1	6,6	1			6,14	2
3408·7	2	2,3	1	3271·6	2	5,13	2
3402·7	3	1,2	1	3269·3	4	9,0	1
3399·7	1	4,4	1	3268·2	1	4,12	2
3397·7	6	0,1	1	3266·9	1	8,15	2
3393·8	2	3,3	1	3265·8	2	3,11	2
3389·7	3	2,2	1	3264·2	3	11,1	1
3384·7	8	1,1	1	3261·7	1	2,10	2
3381·1	1	4,3	1	3259·9	1	6,13	2
3374·3	1	6,4	1	3258·6	1	13,2	1
3368·5	1	5,3	1	3257·9	2	10,0	1
3366·5	6	1,0	1			1,9	2
3358·8	4	3,1	1	3257·0	2	5,12	2
3355·8	1	6,3	1	3253·3	3	12,1	1
3352·9	8	2,0	1			4,11	2
3342·6	1	7,3	1	3248·9	1	7,13	2

Table 1. Ultra-violet absorption bands of Na_2 (*cont.*)

λ_{air} (A.)	Intensity	v',v''	System
3247·9	1	14,2	1
3246·2	2	2,9	2
3242·4	3	13,1	1
		5,11	2
		1,8	2
		4,14	2
3238·3	2	4,10	2
3236·5	2	12,0	1
3235·2	2	3,9	2
3231·8	2	14,1	1
3227·4*	2	16,2	1
		5,10	2
		1,7	2
3225·6	1	13,0	1
3223·6	2	0,6	2
		4,9	2
3221·7	1	15,1	1
3219·5	2	3,8	2
3217·7	1	17,2	1
3215·5	2	2,7	2
3211·9	1	16,1	1
3207·0	1	18,2	1
3204·8	2	3,7	2
3202·1	1	17,1	1
3200·0	2	2,6	2
3196·7	3	19,2	1
		1,5	2
3191·5	1	21,3	1
3187·8	1	20,2	1
3181·4	2	1,4	2
3178·2	2	0,3	2
3176·4	1	21,2	1
3170,8	2	2,4	2
3160·0	2	3,4	2
3156·2	2	2,3	2
3151·6	2	1,2	2
3145·2	2	3,3	2
3140·0	2	2,2	2
3135·7	3	1,1	2
3131·2	2	0,0	2
3125·1	2	2,1	2
3120·5	2	1,0	2
3118·0	2	1,8	3
3110·6	1	3,9	3
3109·5	2	2,0	2
3107·4	1	2,8	3
3103·4	1	4,1	2
3098·6	1	3,0	2
3097·0	3	6,2	2
		3,8	3
3092·9	3	5,1	2
		2,7	3
3089·8	3	5,9	3
3088·2	3	4,0	2
		1,6	3
3086·6	2	6,2	2
		13,6	2
		4,8	3
3082·2	3	6,1	2
		3,7	3
3078·0	3	5,0	2
		3,6	3
3076·3	4	8,2	2
		11,4	2
		5,8	3
3074·1	3	1,5	3
3072·6	1	13,5	2
3071·6*	2	7,1	2
		10,3	2
		4,7	3
3069·6	1	7,9	3
3067·6	3	3,6	3
		6,0	2
		12,4	2
3060·1	4	1,4	3
3057·7*	2	13,4	2
		7,0	2
3056·6	2	10,2	2
3055·6	4	0,3	3
3054·5	1	12,5	2
3052·8	1	12,3	2
3052·2	2	9,1	2
3049·0	1	14,4	2
		2,4	3
3047·6	4	8,0	2
		11,2	2
3046·0	1	16,5	2
3045·6	4	1,3	3
3043·6	1	13,3	2
3041·7	1	0,7	4
3041·0	4	0,2	3
3039·0	1	3,9	4
3034·8	3	14,3	2
		2,3	3
3031·7	3	1,7	4
3031·0	5	1,2	3
3027·9	3	0,6	4
3027·2	5	0,1	3
3026·0	2	3,3	3
3021·0	2	2,2	3
3018·5	3	1,6	4
3016·9	5	1,1	3
3014·7	2	0,5	4
3012·1	5	0,0	3
3009·0	4	2,6	4
3007·0	4	2,1	3
3004·7	2	1,5	4

Table 1. Ultra-violet absorption bands of Na_2 (*cont.*)

λ_{air} (A.)	Intensity	v',v''	System
3002·3	5	1,0	3
3000·3	2	4,0	4
2995·7	4	2,5	4
2992·6	5	2,0	3
2990·9	3	1,4	4
2987·9	4	4,1	3
2986·4	6	0,3	4
2983·1	6	3,0	3
2978·2	4	5,1	3
2977·0	5	1,3	4
2973·8	4	4,0	3
2972·0	4	0,2	4
2968·6	5	6,1	3
2964·6	5	5,0	3
2963·8	4	8,2	3
2959·6	6	7,1	3
2958·6	5	0,1	4
2955·2	6	6,0	3
2953·7	4	9,2	3
2949·7	6	8,1	3
		1,1	4
2948·2	6	2,1	4
2945·5	10	7,0	3
2944·0	5	0,0	4
2941·6	3	12,3	3
2936·3	8	8,0	3
		11,2	3
2935·6	8	1,0	4
2932·5	6	10,1	3
2931·7	3	3,1	4
2928·6	6	12,2	3
2927·6	8	9,0	3
2920·6	5	2,0	4
2924·4	3	11,1	3
2922·9	3	4,1	4
2920·3	5	13,2	3
2919·1	3	10,0	3
2917·2	3	3,0	4
2915·8	5	12,1	3
2914·7	4	5,1	4
2912·9	3	17,4	3
2912·0	3	14,2	3
2908·8	3	4,0	4
2907·4	2	16,3	3
2906·9	3	6,1	4
2904·2	3	18,4	3
2903·5	3	15,2	3
2901·3	1	5,0	4
2899·9	2	17,3	3
2898·8	2	10,3	4
2897·7	2	7,1	4
2896·0*	1	19,4	3
2892·9	1	6,0	4
2891·9*	1	18,3	3
2891·0*	1	15,1	3
2889·7	2	8,1	4
2888·4	1	20,4	3
2886·7	2	17,2	3
2884·5	1	7,0	4
2884·0	1	19,3	3
2882·4	1	12,3	4
2880·1	2	9,1	4
2877·4	2	11,2	4
2876·3	1	8,0	4
2874·0	1	13,3	4
2872·8	1	10,1	4
2869·8	1	12,2	4
2866·5	1	14,3	4
2837	1	0,2	5
2829	1	1,2	5
2824·5	3	0,1	5
2816	1	1,1	5
2808	2	2,1	5
2800	1	3,1	5
2792	2	4,1	5
2780·5	2	4,0	5
2773	2	5,0	5
2765	3	6,0	5
2758	3	7,0	5
2750	5	8,0	5
2746	2	0,7	6
2742	1	9,0	5
2738	2	1,7	6
2735	5	0,6	6
2730	4	2,7	6
2727	3	1,6	6
2719	2	2,6	6
2701	3	3,5	6
2690	2	3,4	6
2672	1	4,3	6
2665	2	5,3	6
2658	1	6,3	6
2654	1	5,2	6
2651	1	7,3	6
2647	1	6,2	6
2643·5	1	5,1	6
2640·5	1	7,2	6
2637	1	6,1	6
2630	1	7,1	6
2623	1	8 1	6
2617	1	9,1	6
2612	1	8,0	6
2605	1	9,0	6
2598·5	1	10,0	6
2577	1	0,4	7
2570	1	1,4	7
2563*	1	2,4	7
2552*	1	2,3	7
2535*	1	3,2	7
2528*	1	4,2	7
2519*	1	4,1	7
2503*	1	5,0	7
2496*	1	6,0	7

Table 2. Na_2 ultra-violet bands in emission

λ_{air} (A.)	Intensity	v',v''	System	λ_{air} (A.)	Intensity	v',v''	System
3366	2	1,0	1	3190	1		
3358	4	3,1	1	3187	1	20,2	1
3353	4	2,0	1	3179	1		
3346	4	4,1	1				
3341	5	3,0	1	3073	2	1,5	3
3334	5	5,1	1	3068	1	0,4	3
3328	4	4,0	1	3059	2	1,4	3
3316	4	5,0	1	3045	4	1,3	3
3312	2			3040	4	0,2	3
3309	2	7,1	1	3031	5	1,2	3
3304	2	6,0	1	3026	5	0,1	3
3298	2	8,1	1	3021	5	2,2	3
3292	4	7,0	1	3017	5	1,1	3
3286	4	9,1	1	3012	5	0,0	3
3280	4	8,0	1	3008	5	2,1	3
3275	4	10,1	1	3002	4	1,0	3
3269	3	9,0	1	2997	2	3,1	3
3264	4	11,1	1	2959	2	7,1	3
3253	4	12,1	1	2950	2	8,1	3
3247	2	11,0	1	2941	4	9,1	3
3236	2	12,0	1	2939	4		
3229	2	16,2	1	2936	4	8,0	3
3226	2	13,0	1	2932	2	10,1	3
3221	4	15,1	1	2929	2		
3218	1	17,2	1	2927	4	9,0	3
3215	2	14,0	1	2923	2		
3207	1	18,2	1	2921	2	13,2	3
3202	1	17,1	1	2918	4	10,0	3
3197	1	19,2	1	2908	2		

§ 3. VIBRATIONAL ANALYSIS

The absorption bands measured in the present investigation have been found to belong to seven systems. Of these, only three systems, viz. 1, 2 and 3, appear at all extensive. The remaining four look like fragments, and until they are photographed under more favourable conditions, the genuiness of their separate existence will remain questionable.

The classification and assignment of quantum numbers have already been given in table 1. In addition, tables 3, 4 and 5 give the Deslandres schemes for the important portions of the three well-developed systems. The mean differences for the upper state given in these tables do not appear quite regular. Since the probable errors in measurements are mostly as high as ± 0·3 A. and sometimes even higher than this, the irregularities shown in these differences may be due to experimental errors and do not seem worthy of any special investigation with regard to perturbation or the like phenomena unless the measurements are improved. A somewhat similar irregularity is also shown by the differences for the ground state, which is known to be free from any perturbation. A few bands appear in brackets in these tables; they have not been observed but have been obtained by interpolation, and have been utilized for calculating the differences

Table 3. Vibrational scheme for the first ultra-violet system of Na_2 bands.

	0		1		2		3		4		5		6		7		8		9	Mean diff.
0					29262	155	(29107)	154	28953						28518	147	(28371)	146	28225	
					118		117		117						107		109			114
1	29696	160	29536	156	29380	156	(29224)	154	29070	149	(28921)	149	28772	147	(28625)	145	28480			
	120				113		114		120		111		110		106					113·5
2	29816				29493	155	29338	148	29190	158	29032	150	29882	151	28731					
	110				117		119		106		108		106							111
3	29926	162	29764	154	(29610)	153	29457	161	29296	156	(29140)	152	28988							
	114						110		111		116		117							113·5
4	30040						29567	160	29407	151	(29256)	151	29105	152	28953					
	108						111													109·5
5	30148	162	29986				29678													
	109																			109
6	30257																			
	113																			113
7	30370	158	30212																	
	105		102																	103·5
8	30475	161	30313																	
	104		108																	106
9	30579	157	(30422)																	
	107		107																	107
10	30686	157	30529																	
	102		98																	100
11	(30788)	161	30627																	
	101		102																	101·5
12	30889	160	30729																	
	104		103																	103·5
13	30993	161	30832																	
			102																	102
14			30934																	
			97																	97
15			31031																	
			94																	94
16			31125																	
Mean diff.		160		155		155		155·5		153·5		150·5		150		146		146		
Loomis & Nusbaum		158		(156)		(155)		(153·5)		(151)		(150·5)		(149)		(146·5)		(146)		

Table 4. Vibrational scheme for the second ultra-violet system fo Na_2 bands

v' \ v''	0		1		2		3		4		5		6		7		8		9		10		11	Mean diff.
0	31927						31453						31012											
	110												113											111·5
1	32037	155	31882	161	31721				31424	151	31273	148	31125	149	30976	144	30832	146	30686					
	113		108		115				105				116		114		110		110					111
2	32150	160	31990	154	31836	161	31675	146	31529				31241	151	31090	148	(30942)	146	30796	146	30650			
	112						110		98 (?)						104		110		105					108
3	32262						31785	158	31627	154	31473				31194	142	31052	151	30901					
	110																		111					110·5
4	32372	159	32213																31012	140	30872	143	30729	
	107		110																		104		103	106
5	32479	156	32323																		30976	144	30832	
	110		112																					111
6	32589	154	32435	155	32280																			
	106				108																			107
7	32695				32388																			
	108				109																			108·5
8	32803	153	(32650)	153	32497																			
			104		105																			104·5
9			32754	152	(32602)																			
					105																			105
10					32707	160	32547																	
					96		100																	98
11					32803	156	(32647)	150	32497	159	32536	148	32388											
							100		92		97													96
12							32747	158	32589	155	(32633)													
							99		106		96													102·5
13							32846	151	32695		32729													
							96		93		91													95
14							32942	154	32788		32820													
																								96
15																								
																								91
16																								
Mean diff.		156		155		159		153		155		148		150		145		148		143		143·5		
Loomis & [illegible]		(158)		(156)		(155)		(153·5)		(151)		(150·5)		(149)		(146·5)		(146)		(144)		(143)		

Table 5. Vibrational scheme for the third ultra-violet system of Na_2 bands

v' \ v''	0		1		2		3		4		5		6		7		8		9	Mean diff.
0	33190	166	33024	150	32874	157	32717													
	108		113		109		107													109
1	33298	161	33137	154	32983	159	32824	155	32669	149	32520	148	32372				32062			
	108		110		109								107				109			109
2	33406	159	33247	155	33092								32479	156	32323	151	32172			
	106												110		112		110			109·5
3	33512												32589	154	32435	153	32282	149	32133	
	105												106		112					108
4	33617	158	33459										32695	148	32547					
	105		109																	107
5	33722	154	33568														32497	142	32355	
	107		108																	107·5
6	33829	153	33676																	
	111		103																	107
7	33940	161	33779																	
	106		112																	109
8	34046	155	33891																	
	102																			102
9	34148																			
	99																			99
10	34247	156	34091																	
			98																	98
11			(34189)																	
			97																	97
12			34286	150	34136															
			99		97															98
13			34385	152	34233															
			99		98															98·5
14			34486	155	34331															
			100		100															100
15			34586	155	34431															
			97		101															99
16			34683	151	(34532)	147	34385													
					100		91													95·5
17					34632	156	34476													
							93													93
18							34569													
Mean diff.		158		153		155		155		149		148		153		152		145·5		
Loomis & Nusbaum		(158)		(156)		(155)		(153·5)		(151)		(150·5)		(149)		(146·5)		(146)		

wherever only few bands for this purpose are available. The numbers given in brackets below the mean differences for the ground state are the values of the latter obtained from the work of Loomis and Nusbaum (1932). Further, the bands lie along a parabola whose shape is similar to what we can expect for the relative values of w_v'' and w_v' of the states involved in these systems.

Weizel and Kulp (1930), who have analysed the bands measured by Walter and Barratt (1928), consider the bands between λ 3370 A. and λ 3180 A. to belong to as many as three systems, while Kimura and Uchida (1932) consider the bands in the same region to belong to two systems. In the present scheme, however, all the intense bands in this region have been placed under system 1 only. It is also possible to accommodate all the bands due to Walter and Barratt in this region into a single system which corresponds to system 1 of the present arrangement. The reason why Weizel and Kulp considered them to belong to more than one system is that a few bands near λ 3303 A. are absent from the spectrogram, having been masked due to continuous absorption resulting from the broadening of the principal-series line. In the same way, systems 4 and 5 of Weizel and Kulp seem to be really another single system corresponding to system 3 of the present classification. System 4 of theirs corresponds to the right limb and system 5 to the left limb of the Condon parabola of the present system 3.

The classification of the emission bands is given in table 2. Bands between λ 3370 A. and λ 3180 A. correspond to the first u.v. system observed in absorption. The emission bands correspond to strong ones on the left limb of the Condon parabola. Bands between λ 3070 A. and λ 2900 A. correspond to the third system in absorption. Only the intense bands have appeared in emission. The second system observed in absorption, which is weaker than the first or the third system, has not appeared in emission. Bands between λ 3370 A. and λ 3600 A. (reported by Kimura and Uchida) are also absent.

The vibrational constants and heats of dissociation for the upper states of the three well developed systems are given in table 6. The heat of dissociation has been calculated by Birge and Sponer's extrapolation method, and its value thus suffers from the defects accompanying this method, especially when the range of extrapolation is large. The calculations for the dissociation products of these states shown in the last column of table 6 are given in the next section.

Table 6. Molecular constants for the upper states of the first three ultra-violet systems of Na_2 bands

Upper-state constants	$\nu_{0,0}$ (cm.$^{-1}$)	w_0' (cm.$^{-1}$)	$x_0'w_0'$ (cm.$^{-1}$)	D' (cm.$^{-1}$)	ν_{atom} (cm.$^{-1}$)	Dissociation products
System 1	29585	115	0·6	5500	28930	$3\ ^2S + 4\ ^2P$ or $3\ ^2S + 3\ ^2D$
System 2	31930	111	0·7	4400	30170	$3\ ^2S + 4\ ^2P$
System 3	33190	109	0·5	5900	32530	$3\ ^2S + 5\ ^2S$

§4. DISSOCIATION PRODUCTS

$\nu_{\text{atom}} = \nu_{0,0} + D' - D''$, where the terms used have their usual significance. Using values of $\nu_{0,0}$ and D' given in table 6, and assuming $D'' = 6160$ cm.$^{-1}$ Loomis and Nusbaum, 1932), we get the following values of ν_{atom}:

$$\nu_{\text{atom}}\,(\text{System } 1) = 28930 \text{ cm.}^{-1},$$
$$\nu_{\text{atom}}\,(\text{System } 2) = 30170 \text{ cm.}^{-1},$$
$$\nu_{\text{atom}}\,(\text{System } 3) = 32930 \text{ cm.}^{-1}.$$

Further, from the line-spectra data (Fowler's *Report*, 1922) it is known that for sodium

$$3\,^2S - 3\,^2D = 29160 \text{ cm.}^{-1},$$
$$3\,^2S - 4\,^2P = 30270 \text{ cm.}^{-1},$$
$$3\,^2S - 5\,^2S = 33200 \text{ cm.}^{-1}.$$

A comparison would thus indicate that the upper states of the first three systems of ultra-violet bands dissociate into a $3\,^2S$ and an excited $3\,^2D$, $4\,^2P$ and $5\,^2S$ atoms. We should then, however, expect two more systems of bands approximately in the same region as these systems, because each of the combination $3\,^2S$ and $3\,^2D$ or $3\,^2S$ and $4\,^2P$ is theoretically capable of giving rise to as many as four stable states, transitions to two of which from the $^1\Sigma_g{}^+$ ground state are permissible. There is no evidence for these systems, nor has any system been observed whose upper-state dissociation products could be $3\,^2S + 4\,^2S$ atoms of sodium. An equally likely dissociation product for the first system could be $3\,^2S + 4\,^2P$ atoms, for the discrepancy of about 1000 cm.$^{-1}$ would not be beyond the range of probable errors involved in calculating D'. Both the assignments for this state are therefore given in table 6. The assignments should, however, be considered merely tentative. These can be settled only if bands can be photographed at still higher dispersion, which might reveal members of higher v' values and also give the structure. Further work on this is in progress.

§5. ACKNOWLEDGMENTS

In conclusion, I wish to express my gratefulness to Assistant Professor R. W. B. Pearse, D.Sc., for his kind help and suggestions and to Professor D. K. Bhattacharya for his ungrudging help in the experimental work in absorption. My thanks are also due to the University of Patna, India, for the award of a scholarship which has enabled me to continue the work here.

REFERENCES

BHATTACHARYA, D. K. and SINHA, S. P., 1943. *Ind. J. Phys.*, **17**, 131.
FOWLER, A., 1922. *Report on Series in Line Spectra* (London : Physical Society), p. 99.
FREDRICKSON, W. R., 1929. *Phys. Rev.*, **34**, 207.
FREDRICKSON, W. R. and STANNARD, C. R., 1933. *Phys. Rev.*, **44**, 633.
FREDRICKSON, W. R. and WATSON, W. W., 1927. *Phys. Rev.*, **30**, 429.
KIMURA, M. and UCHIDA, Y., 1932. *Sci. Pap. Inst. Phys. Chem. Res. (Tokio)*, **18**, 109.
LOOMIS, F. W. and NUSBAUM, R. E., 1932. *Phys. Rev.*, **40**, 380.
LOOMIS, F. W. and WOOD, R. W., 1928. *Phys. Rev.*, **32**, 223.
WALTER, J. M. and BARRATT, S., 1928. *Proc. Roy. Soc.*, A, **119**, 265.
WEIZEL, W. and KULP, M., 1930. *Ann. Phys., Lpz.*, **4**, 971.
WOOD, R. W., 1909. *Phil. Mag.*, **18**, 530.
WOOD, R. W. and GALT, R. H., 1911. *Astrophys. J.*, **33**, 72.

EXPERIMENTS IN MULTIPLE-GAP LINEAR ACCELERATION OF ELECTRONS

By W. D. ALLEN and J. L. SYMONDS,
Radiophysics Laboratory, Sydney, N.S.W.

MS. received 6 *December* 1946

ABSTRACT. Using a CV76 magnetron at a wave-length of 10·0 cm. (maximum power 500 kw.; power available for acceleration, 300 kw.), and a 3-stage single cavity, electrons were accelerated to a voltage of 0·85 Mev. The expected figure was approximately 1·1 Mev. The paper describes the R.F. work involved in obtaining the result.

§ 1. INTRODUCTION

The pioneer work on the linear accelerator of Lawrence and Sloan (1931) and Sloan (1935), indicated the possibility of obtaining fast particles with relatively low H.F. voltages, at frequencies of the order of 30 Mc./s. The linear accelerator has since been eclipsed by the development of the cyclotron: but its possibilities have again become prominent by the development of valves which furnish high-pulsed power at centimetre wave-lengths.

The use of 1200 Mc./s. power in a single cavity has been developed in this Laboratory and described (Bowen, Pulley and Gooden, 1946). To make optimum use of the power available, however, a series of, say, N gaps is necessary: since the power is then distributed between the gaps, the voltages per gap drops by N and the overall voltage increases by $\sqrt{N}$. There are various ways of doing this. A TM_{01} mode can be loaded with irises so that the phase velocity down the guide is equal to the particle (electron) velocity; so that the electron is carried forward continuously on the crest of the wave. The power can be fed into a series of re-entrant cavities, each fed separately, with suitable phasing from an arterial guide. Or the power can be fed into a single cavity, designed so that the voltages at consecutive gaps are suitably reversed. The last method was the one adopted in this work. It had the advantages that good bunching was achievable between first and second gaps: that by virtue of the tight coupling between separate sections, no drift of R.F. phase between the gaps was possible: and that the cavity acted as its own frequency stabilizer.

In what follows, a description is given of the cavity design, of phasing considerations, of high power work and of the acceleration experiments.

§ 2. CAVITY DESIGN

The essential features of the cavity are exhibited in figure 1 A. It consists in effect of a series of re-entrant cavities (figure 1 B) placed back to back, with a cut along the periphery of the dividing walls to provide the electrical coupling between the sections. At each end there is a half-section, in the one case for convenience of feeding and in the other of electron injection. The cut in the dividing wall necessitates the support of the centre " baffles "; this is accomplished by quarter-wave stub supports, as shown in the section view of figure 1 A.

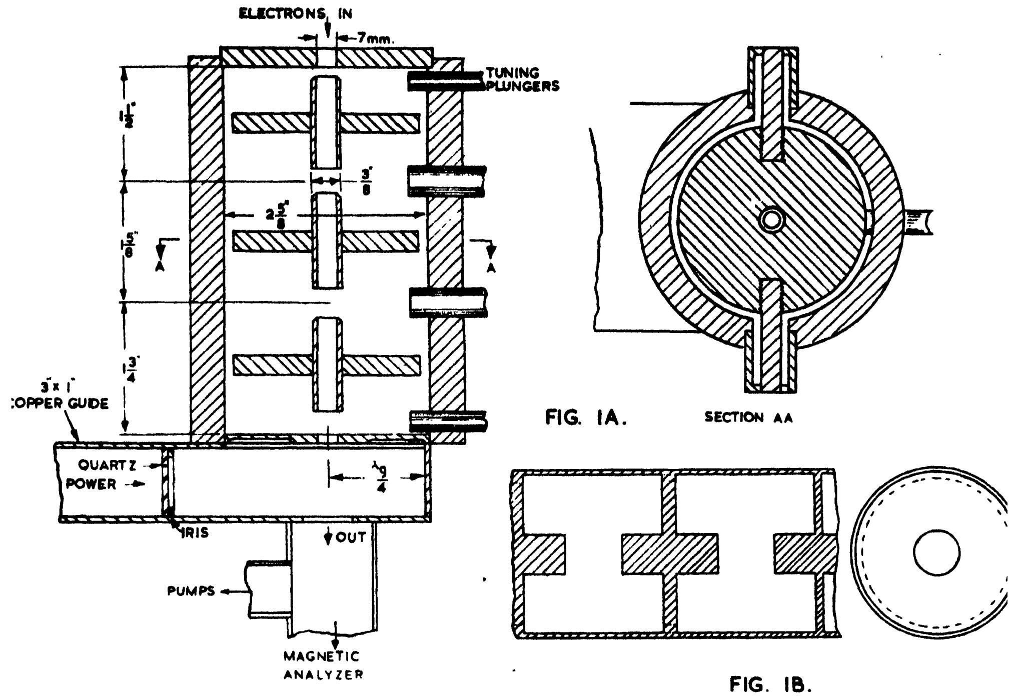


FIG. 1A.

FIG. 1B.

Figures 1 A and 1 B.

The experiments on individual sections of the cavity were carried out on a cavity of the type shown in figure 2.1. This cavity may be regarded as two conventional re-entrant cavities, with gaps at B and C, coupled together by the annulus between disc and chamber wall at A. If we make the customary approximate representation of the re-entrant cavities as a resonant circuit

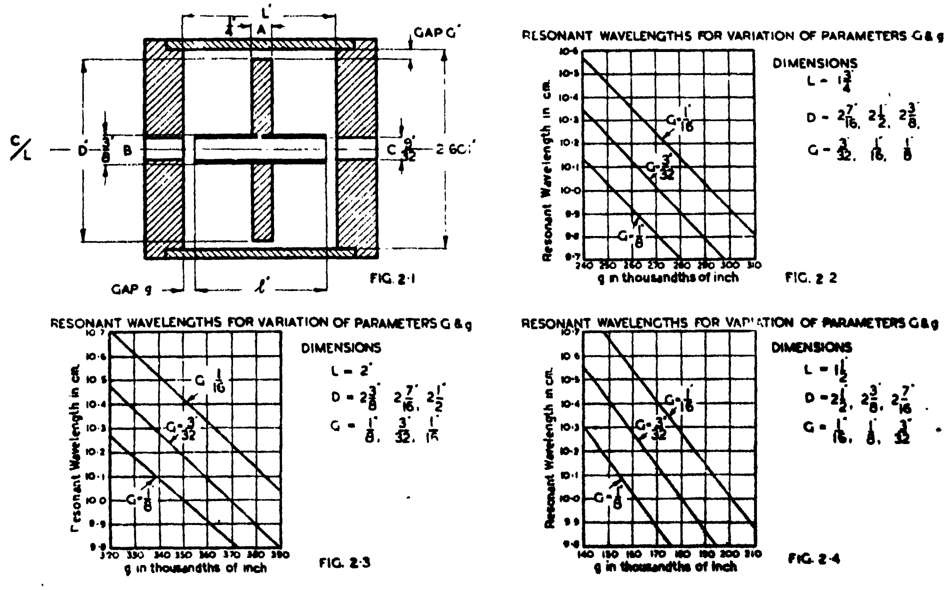


FIG. 2·1

FIG. 2·2

FIG. 2·3

FIG. 2·4

Figures 2.1-2.4.

L and C_1 (figure 3), then we have two such circuits coupled by the capacity C_2 at A. Such a circuit has one resonant frequency with zero voltage at A, with voltages at B and C of opposite sign: this was called the "fundamental" mode. In the other, the desired mode, the voltages at B and C are of the same sign, being opposite to the voltage at A. By virtue of the fact that the gaps of the cavity of figure 2.1 face in opposite directions, however, the instantaneous voltages across the two gaps B and C are always in opposite sense. The phases at A, B and C for both modes were confirmed by measurement with a phase meter.

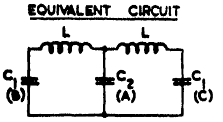


Figure 3.

The dimensions of the cavity (figure 2.1) were determined by the following considerations:—

(a) *Size of inner tube.* This was required to be large, to transmit a reasonable current of electrons; but small, to give maximum shunt resistance; 3/8″ O.D., 9/32″ I.D., was selected as the reasonable compromise.

(b) *Disc thickness and external wall diameter.* To give adequate conduction of the heat generated in the central baffles, the disc was chosen to be $\frac{1}{4}$″ thick. The gap size was desired small, so as to maximize acceleration over a small portion of the R.F. cycle, yet not so small as to cause field emission from the tube ends. A total gap size of $\frac{1}{2}$″ ($\frac{1}{4}$″ at each end of figure 2.1) and an external wall diameter of $2\frac{5}{8}$″ approx. were decided upon. Figures 2.2 and 2.4 show the variation of resonant wave-length of the cavity 2.1 with the variation of various parameters in the cavity.

The assumption in the experiments was that the voltage at A, figure 2.1, would be small compared with the voltage developed at B or C, and that most of the available power would be employed in generating the voltage where it was required. This arose from consideration of the electrostatic capacities at A and B, or from a simple consideration of the relative diameter of disc and tube. That this elementary picture was incorrect was suggested by the following facts:—

(*a*) The fundamental and desired modes differed in frequency by 20%. This would require the ratio C_2/C_1 (figure 3) to be 22%.

(*b*) The resonant frequency sensitivity to variation of gap G (figure 2) is only one-half the resonant frequency sensitivity to variation of gap g.

(*c*) When a polythene rod was inserted so as to fill the tube and gap g, the frequency change was 6%; when a polythene annulus filled gap G, the frequency change was 3·8%.

It is difficult to give an accurate interpretation of these facts: but at least it would seem that the voltage at A is some 20–30% of the voltage at B or C.

The coupling to the cavity was first investigated by a single slot coupled to a simple re-entrant cavity (figure 4). It was found that the reflection coefficient was quite closely proportional to the slot length a over a wide range (30%): this would correspond to the coupling factor being proportional to a^n, where experiment showed $n = 2{\cdot}5$. The theory of coupling given by H. A. Bethe suggests that the coupling of a cavity to a guide by a slot is determined by the magnetic "polarizability" of the slot, which in the case of a narrow elliptical slot is given by

$$\frac{\pi}{3}\left[\frac{a^3}{\log_e(4a/b)-1}\right],$$

where b is the smaller semi-axis. In our case, neglecting the difference between rectangular and oval slots, this would give an effective value of n of 2.6. For the double slot and single cavity as shown in figure 4 the slot length required is 0·7″: for the 3-section cavity, the slot length was 0·95″. Generally speaking, the tolerance on the slot length was 0·02″.

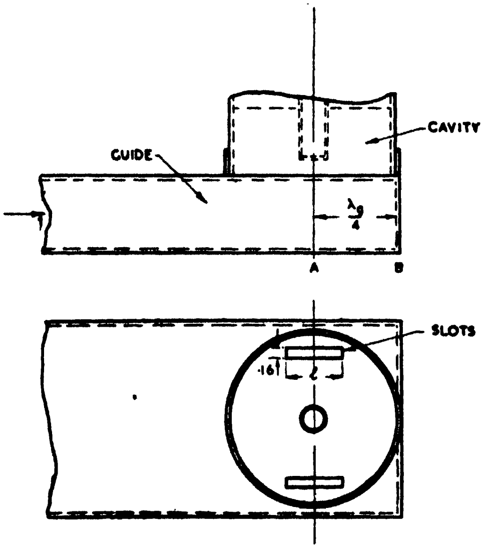


Figure 4.

Tuning of the 3-section cavity was accomplished by the plungers shown in figure 1 A; but as these were also in the region of some electric field, the degree of the tuning that could be achieved by them was only 0·5%. To ensure the correct initial frequency, the discs were at first made oversize, and then turned down until the required wave-length was reached (10·01 cm.). The tuning range was then 10·01–9·96, as required by the magnetron. The vacuum was maintained in the guide by a quartz window waxed to an inductive iris, the system as a whole giving negligible reflection. The 3-section cavity had, in addition to the desired mode (10 cm.) and the fundamental mode (12·1 cm.), two other modes: 10·4 and 11·4 cm. These, however, were all well outside the frequency range of the magnetron. The eventual Q of the cavity, when matched to the guide, was 4000: this reduced, after considerable operation (presumably due to sparking), to 3000.

§ 3. PHASING

The determination of cavity dimensions depends upon the voltage expected across the first few gaps. From the 25 cm. observations, it was considered that a total available peak power of 250 kw. into the cavity and effective shunt resistance of 0·5 MΩ. were reasonable minimum estimates. These figures gave a peak voltage of 300 kv. per gap, or 150 kv. across the first gap. The potential distribution across the gaps was approximated by a method due to R. D. Hill (1945), and step-by-step integration conducted across the gap. In this way, the incident and emergent phases were determined. The inter-gap distances were not critical, and were determined as $1\frac{1}{2}''$, $1\frac{5}{8}''$ and $1\frac{3}{4}''$. The electrons were quite well bunched between first and second gaps; thus, electrons entering the first gaps with phase between 0·4 and 1·2 radian entered the second gap with phase between 0·5 and 1·0 radians.

§ 4. HIGH-POWER WORK

The work on feeding the cavity with power followed the principles laid down by J. R. Pierce and developed in this Laboratory by B. Y. Mills. It can be described with reference to figure 5. The first step is to determine the *Rieke Diagram* of the magnetron, i.e. to plot on the Smith chart the contours of the frequency and power developed by the magnetron when looking at admittances corresponding to points on the chart. The admittance that the magnetron "looked at" was determined by replacing the magnetron by a low power C.W. coaxial standing-wave detector connected to a dummy output seal. This assumes, it is true, that the dummy output seal is identical with the magnetron output seal, which is highly unlikely. But we are mainly concerned with replacing the loads, with which the Rieke is determined, by the cavity, so that, since the method is essentially a substitution method, the difference between magnetron output and dummy output is negligible.

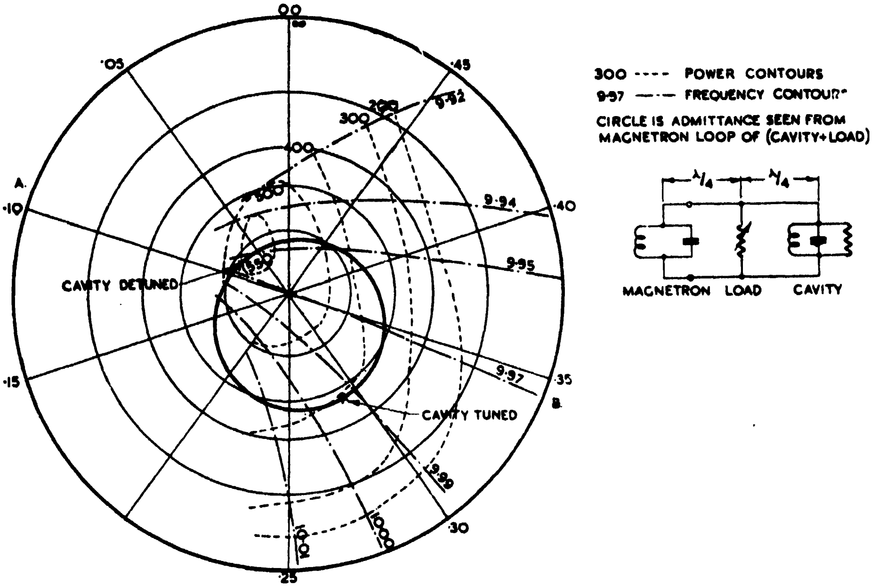


Figure 5. Rieke diagram. cv75. 40A. 2300 gauss.

Figure 5, then, is the Rieke diagram referred to the magnetron output loop. The important point here is the region of the "Frequency sink" (approximately marked "Cavity detuned"); here the frequency contours tend to cross over, and the frequency becomes indeterminate. It was necessary to ascertain this point with some care, for it is at this point that the magnetron circuit (figure 5, inset) can be regarded as being located. Suppose now we have a cavity which when tuned is matched to the magnetron output, and we place the cavity at an even number of half wave-lengths from the magnetron. Then the condition for stable operation is that the admittance change with frequency shall be opposite in direction to the change of the frequency contours which it crosses. The point of match will be a stable point of operation. But the admittance circle of the cavity, with varying frequency, touches the edge of the diagram at A: and by virtue of the crossing contours at the frequency sink, there are also other stable points of operation at points far off the resonant frequency of the cavity, at which the magnetron is developing low power. It is on these modes that the magnetron prefers to operate, if the cavity is half a wave-length from the magnetron. If it is a quarter wave-length, then the point of match is definitely unstable; the magnetron will operate, again, only on points near the detuned point B on the diagram.

The only choice, therefore, is to make the magnetron operate with the cavity half a wave-length away, and yet move the detuned point inside the frequency sink. This is done by inserting a resistive (water) load at the quarter wave point. The resulting admittance circle of (cavity + load) obtained by varying the cavity resonance, is shown by the thick black line of figure 5. As a consequence of the inclusion of the load, a fraction of the power is absorbed by the load, and the magnetron works at a point appreciably off the point of maximum power development; so that, with a 500-kw. magnetron, a maximum of 300 kw. only was eventually available for acceleration of electrons.

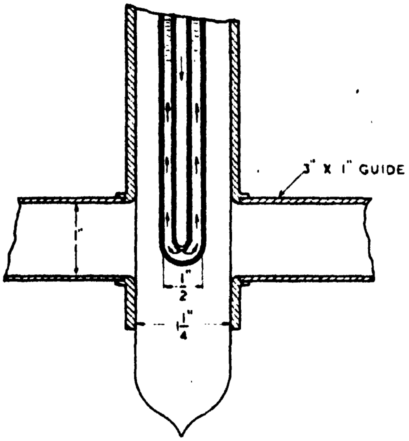


Figure 6 (*a*).

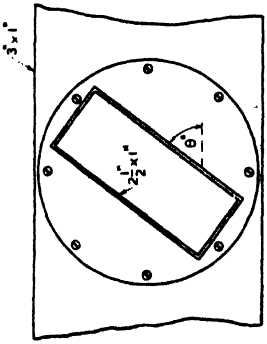


Figure 6 (*b*).

For purposes of experiment, it was desirable to have a high-power load which behaved as a pure variable conductance. Two such loads were developed and are shown in figure 6. The first is simply a water tube which can be pushed in and out of the centre of the wide face of the guide, the coupling varying with the depth of penetration. The tube was ½" diameter, and was carried in a jacket of Freon, to minimize sparking. It behaved very nearly as a pure conductance over the 0·25 to 1·25Y_0 and, in view of the ease of manipulation, was used throughout the experiments. The other, possibly a better load for higher powers, was essentially a T-junction on the broad side of the 3" × 1" guide, the side arm being 2½" × 1" guide terminated with a matched water load. When in the normal position, as an E-plane stub, this load was very nearly Z_0 in series with the arterial 3" × 1" guide: when at right angles to this, the series impedance was negligible. Intermediate positions showed series impedances closely proportional to $\sin^2\theta$, as one would expect.

Observations during operation showed that the ratio of the power in the water load to that in the cavity was considerably higher than that expected from the C.W. measurement.

It was decided that this was due to three effects:—

1. Build-up in the cavity reduced the mean power developed in it.
2. After the magnetron pulse is over, some of the decaying oscillation in the cavity is fed back through the guide to the water load.
3. The main point is that at the beginning of the pulse the current has not built up in the cavity, which is thus mismatched to the guide. The admittance presented to the magnetron is thus effectively the detuned admittance, where the power developed is 500 kw., all of which is absorbed by the water load. The mean power in the load is, therefore, much higher than the instantaneous power at the end of the pulse, which is, of course, the power relevant for acceleration purposes.

§ 5. ACCELERATION EXPERIMENTS

The electrons to be accelerated were provided by an electron gun, which was initially operated at 12 kv. DC. After passing through the cavity and guide (figure 1 A) they passed into a magnetic analyser. Comparison between the magnetic field and applied voltage at 12 kv. showed agreement within 3%, which is probably within the limits of accuracy of the various meters concerned. A feature at both high (800 kv.) and low voltage observations was a "tail", about 1% of the peak current, extending to some 20–30% of the peak field on either side of the peak; it was presumed to be due to reflection at the edges of the defining slit of the analyser.

During the acceleration experiments, the 25 kv. magnetron pulse was applied to the gun, and the current to the slit was initially picked up by the pulse amplifier. The voltage of 800–850 kv. was not materially changed by adjustment of many of the available variables; the peak was quite sharp, although on the low-voltage side there was a considerable number of electrons (order of 10^{-8} amp. mean) provided by emission of one kind or another in the accelerating gaps. When the current was maximized, 3μA. mean was available at the analyser slit, the voltage dropping to some 800 kv. The duty cycle was 2800, so that some 8 mA. peak was reaching the slit; and as the cavity does not reach peak voltage till towards the end of the pulse, and as acceleration takes place only over a fraction of the R.F. cycle, the instantaneous current is an appreciable fraction of 1 ampere. Considering also that the analyser slit was $\frac{3}{8}'' \times \frac{1}{8}''$ located 18″ from the orifice of the cavity, one concludes that either the focusing of cavity and analyser was good, or that the total peak current was considerable.

When the cavity was examined, it was found that considerable sparking had taken place between discs and cavity wall, particularly at the one next to the guide: as the cavity was effectively being evacuated through this annulus, presumably the pressure here at times could become appreciable. There was some evidence of anti-symmetric voltages at the disc, possibly generated by the tuning plungers, which were projecting considerably into the guide.

The figure of 1100 kv. mentioned in the abstract was arrived at as follows:— The observed maximum power in the cavity was 300 kw. peak, and an estimated shunt resistance of 1 MΩ (compared with some measured 25 cm. values) seemed attainable. These figures give, for each stage, an R.M.S. voltage of 310 kv. or a peak voltage of 450 kv.: which for 3 gaps gives a maximum possible of 1350 kv.

Of this, the peak current, as calculation shows, is 90% of the maximum possible: and the remaining deficit between 850 kv, and 1220 kv. is presumably to be ascribed either to sparking at axial or peripheral gaps, or to the appreciable proportion of the available voltage being taken up at the peripheral gap. Unfortunately circumstances prevented an adequate analysis of these alternatives being made.

ACKNOWLEDGMENT

The work was carried out as part of the programme of the Radiophysics Division of C.S.I.R., Australia.

REFERENCES

BETHE, H. A., *M.I.T. Report*, **43**, 22.
BOWEN, PULLEY and GOODEN, 1946. *Nature, Lond.*, **157**, 840.
HILL, R. D., 1945. *J. Sci. Instrum.*, **22**, 221.
LAWRENCE, E. O. and SLOAN, D. H., 1931. *Phys. Rev.*, **38**, 2021.
PIERCE, J. R. B.T.L. MM 43/140/19.
SLOAN, D. H., 1935. *Phys. Rev.*, **47**, 62.

THERMODYNAMIC RELATIONS FOR TWO PHASES CONTAINING TWO COMPONENTS IN EQUILIBRIUM UNDER GENERALIZED STRESS

BY C. GURNEY,
Royal Aircraft Establishment, Farnborough

MS. received 7 *December* 1946

ABSTRACT. The application of thermodynamics to cases of other than hydrostatic pressure is important in connection with the swelling and flow and fracture of solids under generalized stress. In the present paper the methods of Gibbs are applied to the case of two phases containing the same two component substances in equilibrium with each other.

The problem is first considered in its most general form, each phase being under generalized stress and each containing each component. The more particular problem in which one of the components is absent from one of the phases is then considered, and the particular case in which one of the phases is fluid and, therefore, able to withstand only hydrostatic pressure, is dealt with in some detail. The cases of a two-component fluid phase in equilibrium with a one-component solid phase and a one-component fluid phase in equilibrium with a two-component solid phase are treated together. These cases correspond respectively to what are often called solution and swelling, although there is no logical reason for this nomenclature. The derivatives of pressure on the fluid phase for changes of temperature and changes of each of the components of generalized stress on the solid phase are given. When suitably interpreted, the same formulae apply to both solution and swelling. Formulae for entropy changes with stress and temperature are also given, and the use of other independent variables such as strain, force, and displacement instead of stress is discussed.

§ 1. INTRODUCTION

THE most usual stress system considered in thermodynamics is hydrostatic pressure. Solids can withstand generalized stress, and increasing interest is being shown in the thermodynamics of stress systems involving other than hydrostatic pressure. This subject finds applications in the swelling of substances such as wood and plastics by water, and in the flow of rocks, where in some cases the flow is attributed to solution of highly stressed parts of the rock and deposition of the dissolved material at stress-free places; and it is likely to be important in connection with the fracture of brittle materials which are subject to attack by the surrounding medium. Here the increasing severity of the attack with increasing stress is thought to cause the highly stressed material at the ends of cracks to be preferentially attacked, so that the cracks gradually spread and cause delayed fracture. The subject should also be of importance in metallurgy, where internal stresses due to work hardening or anisotropic thermal expansion may decrease the stability of the structure, and lead to weakening or the development of cracks; and phase transitions due to stress may lead to flow and failure.

The application of thermodynamics to stress systems other than hydrostatic pressure has already been discussed by Gibbs (1876) in his original paper on the equilibrium of heterogeneous substances. In the present paper the methods developed by Gibbs are applied to the case of equilibrium between two phases containing the same two components. The subject is first discussed in its most general aspect, each phase containing each component and each subjected to its own system of generalized stress. On account of the need to satisfy the equality of the chemical potential of each component in both phases, it is not possible to vary only two independent variables at a time, and true partial derivatives of the variables cannot be obtained. It is therefore not possible to arrive at results of much generality. If, however, one of the components is absent from one of the phases, it is only necessary to satisfy the conditions of equality of the chemical potentials of the component common to both phases. For this case, true partial derivatives can be obtained. Two cases of this sort are of practical importance:—(1) A solid two-component phase in equilibrium with a fluid one-component phase: the absorption of water by wood is an example of this sort. (2) The other case is that of a one-component solid phase in equilibrium with a two-component fluid phase: saturated solutions of many salts provide examples of this. For these two cases the partial derivatives of the pressure on the fluid phase with respect to the temperature and with respect to any one of the components of generalized stress acting on the solid phase have been computed. It is of some interest that when suitably interpreted the same formulae apply to both of these cases.

§ 2. THERMODYNAMIC RELATIONS FOR SYSTEM UNDER GENERALIZED STRESS

The stress system acting on a body enters thermodynamics via the work which forces do in moving their points of application. Two of the four thermodynamic energy functions contain "work done" explicitly. These are the energy E and the Helmholtz free energy F. The energy also contains entropy changes explicitly, and is therefore useful when considering rapid changes which may be assumed to take place adiabatically, for such changes take place at constant entropy. We are here more concerned with slow changes which take place isothermally, and we therefore choose F as our energy function because it contains temperature change explicitly. For a phase containing two components, the most general change in F is given by

$$dF = -SdT + dW + \mu_1 dn_1 + \mu_2 dn_2. \qquad \ldots\ldots(1)$$

Here S is entropy, T is temperature, W is work, μ is chemical potential, n is quantity of component. The two components are designated by subscripts 1 and 2.

For hydrostatic pressure the change in work done, dW, is equal to $-pdV$, where V is volume and p is pressure. If we wish to change the independent variable to p, we write $dW = -p\frac{\partial V}{\partial p}dp$. For work done under generalized stress we have similar choice of variables. We may choose the six components of force (three direct forces and three shear forces), and with these it is convenient to choose six components of displacement as associated variables. The components are conveniently distinguished by two numerical suffixes, the first suffix indicating the direction of the normal to the plane on which the force acts and the second suffix the direction of the force. Thus if P and dx are forces and displacements respectively, P_{11}, P_{22}, P_{33}, dx_{11}, dx_{22}, dx_{33} are direct forces and extensional displacements, while P_{12}, P_{23}, P_{31}, dx_{12}, dx_{23}, dx_{31} are shear forces and shear displacements. It is convenient to adopt the shorthand notations P_{ij} and dx_{ij} for force and displacement, where i and j are understood to stand for numbers between 1 and 3. Tension is taken as positive. The work done for a general small change in displacements is then

$$dW = \Sigma_{ij}P_{ij}dx_{ij}. \qquad \ldots\ldots(2)$$

If it is desired to have the P_{ij} as the independent variables we may write for constant temperature and quantities of each component

$$dx_{ij} = \Sigma_{kl}\frac{\partial x_{ij}}{\partial P_{kl}}dP_{kl} \qquad \ldots\ldots(3)$$

and

$$dW = \Sigma_{ij}P_{ij}\Sigma_{kl}\frac{\partial x_{ij}}{\partial P_{kl}}dP_{kl}. \qquad \ldots\ldots(4)$$

Instead of force and displacement it is more usual to use stress and strain as variables. These may be denoted by X_{ij} and e_{ij}. In terms of these variables, change in work done becomes

$$dW = V\Sigma_{ij}X_{ij}de_{ij}, \qquad \ldots\ldots(5)$$

where V is the volume. If we wish to have stress as independent variable, the change in W becomes

$$dW = V\Sigma_{ij}X_{ij}\Sigma_{kl}\frac{\partial e_{ij}}{\partial X_{kl}}dX_{kl}. \qquad \ldots\ldots(6)$$

In the theory of elasticity, strain is the fractional change of length due to stress, but here we use strain to denote proportional change in length due to any cause. It will therefore include change in length due to change in temperature, due to change in composition and due to change in quantity of the phase. The most general expression for dW is, therefore,

$$dW = V\Sigma_{ij}X_{ij}\left(\Sigma_{kl}\frac{\partial e_{ij}}{\partial X_{kl}}dX_{kl} + \frac{\partial e_{ij}}{\partial T}dT + \frac{\partial e_{ij}}{\partial n_1}dn_1 + \frac{\partial e_{ij}}{\partial n_2}dn_2\right). \qquad \ldots\ldots(7)$$

Substituting this value in expression (1) gives

$$dF = \left(-S + V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial T}\right)dT + V\Sigma X_{ij}\Sigma\frac{\partial e_{ij}}{\partial X_{kl}}dX_{kl} + \left(\mu_1 + V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial n_1}\right)dn_1$$
$$+ \left(\mu_2 + V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial n_2}\right)dn_2. \qquad \ldots\ldots(8)$$

For equilibrium between two phases each containing two components, the following relations must hold (Gibbs, 1876):

$$T^{\alpha}=T^{\beta};\quad \mu_1^{\alpha}=\mu_1^{\beta};\quad \mu_2^{\alpha}=\mu_2^{\beta}, \qquad \ldots\ldots(9)$$

where the superscripts α and β designate the phases and the suffixes 1 and 2 designate the components. After any changes in the variables of the system these relations must continue to be valid. Suppose each of the phases is under generalized stress and that we vary one stress denoted by X_{kl} of the phase α and another stress denoted by X_{mn} of the phase β. Then we have

$$\frac{\partial \mu_1}{\partial X_{kl}}\,dX^{\alpha}_{k}=\frac{\partial \mu_1}{\partial X_{mn}}\,dX^{\beta}_{mn}, \qquad \ldots\ldots(10)$$

$$\frac{\partial \mu_2}{\partial X_{kl}}\,dX^{\alpha}_{kl}=\frac{\partial \mu_2}{\partial X_{mn}}\,dX^{\beta}_{mn}. \qquad \ldots\ldots(11)$$

Obviously in general these equations cannot simultaneously be satisfied, and we are therefore not free to vary only one variable of each phase at a time. Some other variable of one of the phases, such as another stress or temperature or quantity of the components, must be varied. The partial change of a variable of one phase cannot therefore be expressed with respect to a variable of the other phase, all other variables remaining constant. In the general case, therefore, results of sufficient generality to justify their inclusion in this paper cannot be obtained.

If, however, one of the phases contains only one component, say component 1, it is only necessary to ensure equality of the chemical potentials of that component. This is a case of common occurrence in practice. It includes, for example, the equilibrium of wood containing absorbed moisture with water, and the equilibrium of many salts with their saturated solutions; and it includes many examples found in metallurgy. We then have for the partial variation in stresses on the two phases

$$\frac{\partial X^{\alpha}_{kl}}{\partial X^{\beta}_{mn}}=\frac{\dfrac{\partial \mu_1^{\beta}}{\partial X_{mn}}}{\dfrac{\partial \mu_1^{\alpha}}{\partial X_{kl}}}. \qquad \ldots\ldots(12)$$

We may evaluate this expression in terms of other variables of the system by using the fact that dF is a total differential and, therefore, that the order of differentiation is of no consequence. Differentiating the third coefficient of expression (8) with respect to stress and the second with respect to quantity of component 1 gives

$$\frac{\partial \mu_1}{\partial X_{kl}}=\frac{\partial}{\partial n_1}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial X_{kl}}\right)-\frac{\partial}{\partial X_{kl}}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial n_1}\right). \qquad \ldots\ldots(13)$$

Expression (12) therefore becomes

$$\frac{\partial X^{\alpha}_{kl}}{\partial X^{\beta}_{mn}}=\frac{\left[\dfrac{\partial}{\partial n_1}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial X_{mn}}\right)-\dfrac{\partial}{\partial X_{mn}}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial n_1}\right)\right]^{\beta}}{\left[\dfrac{\partial}{\partial n_1}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial X_{kl}}\right)-\dfrac{\partial}{\partial X_{kl}}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial n_1}\right)\right]^{\alpha}}. \qquad \ldots\ldots(14)$$

Instead of varying the stress on both phases we may obtain equilibrium by varying two stress components on one phase, so that the resultant change in μ is zero. We would then have an expression similar to expression (14) with α substituted for β and the sign changed.

If, when the stress on the α phase is varied, other stresses are maintained constant and the temperature is varied to maintain equilibrium, we have

$$\left[\frac{\partial\mu_1}{\partial X_{kl}}dX_{kl}+\frac{\partial\mu_1}{\partial T}dT\right]^\alpha=\left[\frac{\partial\mu_1}{\partial T}dT\right]^\beta. \qquad \ldots\ldots(15)$$

Thus

$$\frac{\partial X_{kl}}{\partial T}=\frac{\dfrac{\partial\mu_1^\beta}{\partial T}-\dfrac{\partial\mu_1^\alpha}{\partial T}}{\dfrac{\partial\mu_1^\alpha}{\partial X_{kl}}}. \qquad \ldots\ldots(16)$$

By differentiating the third coefficient of expression (8) with respect to T and the first with respect to n and equating, and using the previously obtained expression for $\frac{\partial\mu_1}{\partial X_{kl}}$, equation (16) becomes

$$\frac{\partial X_{kl}}{\partial T}=\frac{\Delta\left[-\dfrac{\partial}{\partial T}V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial n_1}-\dfrac{\partial S}{\partial n_1}+\dfrac{\partial}{\partial n_1}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial T}\right)\right]}{\left[\dfrac{\partial}{\partial n_1}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial X_{kl}}\right)-\dfrac{\partial}{\partial X_{kl}}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial n_1}\right)\right]^\alpha}. \qquad \ldots\ldots(17)$$

Here the symbol Δ indicates the difference between corresponding quantities for the phases β and α. $T\Delta\frac{\partial S}{\partial n_1}$ is the latent heat of isothermal change of unit quantity of component 1 from a large quantity of the phase α to phase β.

When the stress on one phase is varied, we may also obtain equilibrium by varying the composition of either phase, the other variables being kept constant. Thus when the composition of the other phase is varied

$$\left[\frac{\partial\mu_1}{\partial X_{kl}}\right]^\alpha dX_{kl}=\left[\frac{\partial\mu_1}{\partial n_1}dn_1\right]^\beta, \qquad \ldots\ldots(18)$$

giving

$$\frac{[\partial X_{kl}]^\alpha}{[\partial n_1]^\beta}=\frac{\left[\dfrac{\partial\mu}{\partial n_1}\right]^\beta}{\left[\dfrac{\partial\mu_1}{\partial X_{kl}}\right]^\alpha}. \qquad \ldots\ldots(19)$$

This expression is not as useful as those previously obtained, as $\frac{\partial\mu}{\partial n_1}$ cannot be equated to expressions containing derivatives of strain.

Having thus obtained expressions for the partial derivative of stress with respect to other components of stress, temperature and composition, we may obtain expressions for the partial derivatives of any of the independent variables in terms of any other independent variables. For example, we may obtain the partial variation of composition with temperature from the expression

$$\frac{\partial n_1}{\partial T}=-\frac{\dfrac{\partial X_{kl}}{\partial T}}{\dfrac{\partial X_{kl}}{\partial n_1}}. \qquad \ldots\ldots(20)$$

So far we have considered both phases being acted upon by generalized stress. If one phase is a fluid, and can therefore only permanently withstand hydrostatic pressure we could obtain the partial derivatives of pressure with respect to the other variables by varying X_{11}, X_{22}, and X_{33} by equal amounts simultaneously. Thus

$$\frac{\partial\mu}{\partial X_{11}}dX_{11}+\frac{\partial\mu}{\partial X_{22}}dX_{22}+\frac{\partial\mu}{\partial X_{33}}dX_{33}=\frac{\partial\mu}{\partial p}dp \qquad \ldots\ldots(21)$$

if

$$X_{11}=X_{22}=X_{33}=p \quad \text{and} \quad dX_{11}=dX_{22}=dX_{33}=dp.$$

Alternatively we may write for the fluid phase

$$dF=\left[\left(-S-p\frac{\partial V}{\partial T}\right)dT-p\frac{\partial V}{\partial p}dp+\left(\mu_1-p\frac{\partial V}{\partial n_1}\right)dn_1+\left(\mu_2-p\frac{\partial V}{\partial n_2}\right)dn_2\right]^{\mathrm{F}}. \qquad \ldots\ldots(22)$$

and evaluate $\partial\mu/\partial p$ from cross differentiating appropriate terms of expression (22).

Two cases of a solid phase in equilibrium with a fluid phase are of particular interest. The first is a two-component solid phase in equilibrium with a one-component fluid phase: this is a common case in swelling phenomena. The other case is a one-component solid phase in equilibrium with a two-component fluid phase: this corresponds to equilibrium between a pure solid substance and its saturated solution. As changes in chemical potential and quantity only occur for the substance which is common to both phases, we may drop suffixes and write

$$\mu^{\mathrm{S}}=\mu^{\mathrm{F}}, \qquad \ldots\ldots(23)$$

where μ is the chemical potential of the component common to both phases; the superscripts S and F indicate solid and fluid phases. In the same way $\partial\mu/\partial n$ must be understood to mean the derivative of the chemical potential of the component which is common to both phases with respect to the quantity of this component. With this notation, the same general formulae apply to both the particular cases mentioned above. In the next section we discuss these cases in more detail.

§ 3. DERIVATIVES OF PRESSURE FOR SOLID-FLUID EQUILIBRIUM

In this section we discuss a two-component solid phase in equilibrium with a one-component fluid phase, and a one-component solid phase in equilibrium with a two-component fluid phase. The same formulae apply. For brevity we refer to the former as a case of swelling and the latter as solution, although logically either of these terms might be applied to either of the cases considered.

3.1. *Variation of pressure with stress.*

By the method of § 2 we readily obtain for the variation in pressure on the fluid phase with stress on the solid phase

$$\frac{\partial p}{\partial X_{kl}}=\frac{\left[\frac{\partial}{\partial n}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial X_{kl}}\right)-\frac{\partial}{\partial X_{kl}}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial n}\right)\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(24)$$

This may be rewritten

$$\frac{\partial p}{\partial X_{kl}}=\frac{\left[\frac{\partial V}{\partial n}\Sigma X_{ij}\frac{\partial e_{ij}}{\partial X_{kl}}-\frac{\partial V}{\partial X_{kl}}\Sigma X_{ij}\frac{\partial e_{ij}}{\partial n}-V\frac{\partial e_{kl}}{\partial n}\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}\,.\qquad\ldots\ldots(25)$$

The terms such as $\frac{\partial V}{\partial n}, \frac{\partial e}{\partial X}$ etc. in this and subsequent equations are all functions of the independent variables.

It is of interest to discuss the physical significance of the terms in equation (25). In the case of swelling, $\frac{\partial V^{\mathrm{S}}}{\partial n}$ is the volume swelling of the solid per unit mass of absorbed fluid. $\frac{\partial e_{ij}}{\partial X_{kl}}$ are the reciprocals of the elastic constants of the solid, such as $\frac{1}{E}$ or $-\frac{\nu}{E}$. $\frac{\partial V}{\partial X_{kl}}$ is the volume change due to change in the particular stress under consideration. $\frac{\partial e_{ij}}{\partial n}$ are the changes in strain due to swelling, and $\frac{\partial e_{kl}}{\partial n}$ are the changes (due to swelling) in the strain corresponding to the particular stress which is varied. $\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}$ is the specific volume of the fluid.

In the case of solution some of the terms have very different significance. In this case $\frac{\partial V^{\mathrm{S}}}{\partial n}$ is the specific volume of the solid, and $\frac{\partial e_{ij}}{\partial n}$ and $\frac{\partial e_{kl}}{\partial n}$ are the changes in proportional dimensions due to the solid being dissolved away. $\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}$ is the volume swelling of the fluid per unit mass of dissolved solid.

It is of interest to consider some simple cases:—

Case 1. *Simple direct stress*

$$X_{22}=X_{33}=X_{12}=X_{23}=X_{31}=0.$$

$$\frac{\partial p}{\partial X_{11}}=\frac{\left[\frac{\partial V}{\partial n}X_{11}\frac{\partial e_{11}}{\partial X_{11}}-\frac{\partial V}{\partial X_{11}}X_{11}\frac{\partial e_{11}}{\partial n}-V\frac{\partial e_{11}}{\partial n}\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}\,.\qquad\ldots\ldots(26)$$

In the case of swelling of an isotropic material with shear modulus G and linear strain e and specific volume of pure fluid $\bar{V}$,

$$\frac{\partial p}{\partial X_{11}}=\frac{\left[V\frac{\partial e}{\partial n}\left(\frac{X_{11}}{G}-1\right)\right]^{\mathrm{S}}}{[\bar{V}]^{\mathrm{F}}}\,.\qquad\ldots\ldots(27)$$

In the case of solution off the face on which X_{11} acts,

$$\frac{\partial p}{\partial X_{11}}=\frac{\left[\bar{V}\left(\frac{2\nu X_{11}}{E}-1\right)\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}},\qquad\ldots\ldots(28)$$

where $\bar{V}$ is the specific volume of pure solid. If solution takes place off a stress-free face then $\frac{\partial e_{11}}{\partial n} = 0$ and expression (26) becomes

$$\frac{\partial p}{\partial X_{11}} = \frac{\left[\bar{V}\frac{X_{11}}{E}\right]^{S}}{\left[\frac{\partial V}{\partial n}\right]^{F}}. \qquad \ldots\ldots(29)$$

For stressses small compared with the elastic modulus, expression (28) usually has the biggest value, expression (27) the next biggest, and expression (29) is smallest. The latter is zero at zero stress. At low stresses, increase in compression produces effects of opposite sign to increase in tension for the cases represented by equations (27) and (28). The effect represented by equation (29) is independent of the sign of the stress.

Case 2. *Simple shear stress*

$$X_{11} = X_{22} = X_{33} = X_{23} = X_{31} = 0.$$

Equation (25) gives

$$\frac{\partial p}{\partial X_{12}} = \frac{\left[\frac{\partial V}{\partial n} X_{12} \frac{\partial e_{12}}{\partial X_{12}} - \frac{\partial V}{\partial X_{12}} X_{12} \frac{\partial e_{12}}{\partial n} - V\frac{\partial e_{12}}{\partial n}\right]^{S}}{\left[\frac{\partial V}{\partial n}\right]^{F}} \qquad \ldots\ldots(30)$$

In the case of swelling of an isotropic material, the value of $\frac{\partial e_{12}}{\partial n}$ depends on the rate of change of shear modulus with moisture content and equals $-\frac{X_{12}}{G}\frac{\partial G}{\partial n}$. The middle term is then negligible, and if e is the linear swelling

$$\frac{\partial p}{\partial X_{12}} = \frac{\left[\frac{V}{G} X_{12}\left(3\frac{\partial e}{\partial n} + \frac{1}{G}\frac{\partial G}{\partial n}\right)\right]^{S}}{[\bar{V}]^{F}} \qquad \ldots\ldots(31)$$

In the case of solution off any face, $\frac{\partial e_{12}}{\partial n}$ is zero and (30) becomes

$$\frac{\partial p}{\partial X_{12}} = \frac{\left[\bar{V}\frac{X_{12}}{G}\right]^{S}}{\left[\frac{\partial V}{\partial n}\right]^{F}}. \qquad \ldots\ldots(32)$$

Case 3. *Three unequal principal stresses*

$$X_{12} = X_{23} = X_{31} = 0.$$

This is only discussed for the case of swelling. Solubility would be different on each of the three pairs of opposite faces of a cube, and to obtain equilibrium the liquid in contact with each opposite pair would have to be isolated from that in contact with other pairs. Solubility for this case is not, therefore, further discussed. For swelling

$$\frac{\partial p}{\partial X_{11}} = \frac{V^{\mathrm{S}}\left[X_{11}\left\{-\frac{\partial e_{11}}{\partial n}\left(\frac{\partial e_{22}}{\partial X_{11}}+\frac{\partial e_{33}}{\partial X_{11}}\right)+\frac{\partial e_{11}}{\partial X_{11}}\left(\frac{\partial e_{22}}{\partial n}+\frac{\partial e_{33}}{\partial n}\right)\right\}+\begin{array}{c}\text{similar terms}\\ \text{in } X_{22} \text{ and } X_{33}\end{array}-\frac{\partial e_{11}}{\partial n}\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(33)$$

For an isotropic material

$$\frac{\partial p}{\partial X_{11}} = \frac{\left[V\frac{\partial e}{\partial n}\left\{\frac{2X_{11}-\overline{X_{22}+X_{33}}}{2G}-1\right\}\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(34)$$

The total change due to changes in the three stresses is

$$dp = \frac{\partial p}{\partial X_{11}}dX_{11} + \frac{\partial p}{\partial X_{22}}dX_{22} + \frac{\partial p}{\partial X_{33}}dX_{33}. \qquad \ldots\ldots(35)$$

If in equation (33) we neglect all the terms in the numerator except $-\frac{V\partial e_{11}}{\partial n}$, equation (35) becomes

$$\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}} dp = \left[-V\left\{\frac{\partial e_{11}}{\partial n}dX_{11} + \frac{\partial e_{22}}{\partial n}dX_{22} + \frac{\partial e_{33}}{\partial n}dX_{33}\right\}\right]^{\mathrm{S}}. \qquad \ldots\ldots(36)$$

This is Barkas's equation (12) in his 1945 paper. It applies so long as the stresses are small compared with the shear modulus or to an isotropic material if the initial state is one of hydrostatic pressure. This conclusion has been reached by Warburton (1946) by a different analysis.

3.2. *Variation of pressure with temperature*

By the method of § 2 we obtain for the variation of the pressure on the fluid phase with the variation of the common temperature of both phases when the solid phase is under generalized stress,

$$\frac{\partial p}{\partial T} = \frac{\left[\frac{\partial V}{\partial n}\Sigma X_{ij}\frac{\partial e_{ij}}{\partial T} - \frac{\partial V}{\partial T}\Sigma X_{ij}\frac{\partial e_{ij}}{\partial n}\right]^{\mathrm{S}} + \frac{\partial}{\partial n}\frac{L}{T}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(37)$$

$\frac{\partial L}{\partial n}$ is the latent heat of transfer of unit quantity of the component common to both phases from a large quantity of the solid phase to a large quantity of the fluid phase.

Case 1. *Simple tension*

$$X_{22} = X_{33} = X_{12} = X_{23} = X_{31} = 0.$$

$$\frac{\partial p}{\partial T} = \frac{\left[\frac{\partial V}{\partial n}X_{11}\frac{\partial e_{11}}{\partial T} - \frac{\partial V}{\partial T}X_{11}\frac{\partial e_{11}}{\partial n}\right]^{\mathrm{S}} + \frac{\partial}{\partial n}\frac{L}{T}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}; \qquad \ldots\ldots(38)$$

for an isotropic material and in the case of swelling,

$$\frac{\partial p}{\partial T}=\frac{\left[\frac{1}{T}\frac{\partial L}{\partial n}\right]}{[V]^{\mathrm{F}}}; \qquad \text{......(39)}$$

for an isotropic material and in the case of solution off the stressed faces

$$\frac{\partial p}{\partial T}=\frac{\left[-2\bar{V}\frac{\partial e}{\partial T}X_{11}\right]^{\mathrm{S}}+\frac{1}{T}\frac{\partial L}{\partial n}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}; \qquad \text{......(40)}$$

for an isotropic material and in the case of solution off the stress-free faces,

$$\frac{\partial p}{\partial T}=\frac{\left[\bar{V}\frac{\partial e}{\partial T}X_{11}\right]^{\mathrm{S}}+\frac{1}{T}\frac{\partial L}{\partial n}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \text{......(41)}$$

Case 2. *Simple shear*

$$X_{11}=X_{22}=X_{33}=X_{23}=X_{31}=0.$$

$$\frac{\partial p}{\partial T}=\frac{\left[\frac{\partial V}{\partial n}X_{12}\frac{\partial e_{12}}{\partial T}-\frac{\partial V}{\partial T}X_{12}\frac{\partial e_{12}}{\partial n}\right]^{\mathrm{S}}+\frac{1}{T}\frac{\partial L}{\partial n}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \text{......(42)}$$

If we neglect the variation in shear modulus with temperature and composition, equation (42) becomes for an isotropic material and in the case of swelling

$$\frac{\partial p}{\partial T}=\frac{\frac{1}{T}\frac{\partial L}{\partial n}}{[V]^{\mathrm{F}}}, \qquad \text{......(43)}$$

and in the case of solution off any face

$$\frac{\partial p}{\partial T}=\frac{\frac{1}{T}\frac{\partial L}{\partial n}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \text{......(44)}$$

The major effect of stress system on the value of $\frac{\partial p}{\partial T}$ is, therefore, its effect on the latent heat.

3.3. *Variation of pressure with composition*

Here it will be necessary to treat swelling and solution separately.

For swelling we have

$$\frac{\partial \mu^{\mathrm{S}}}{\partial n}dn=\frac{\partial \mu^{\mathrm{F}}}{\partial p}dp, \qquad \text{......(45)}$$

$$\frac{\partial p}{\partial n}=\frac{\left[\frac{\partial \mu}{\partial n}\right]^{\mathrm{S}}}{[V]^{\mathrm{F}}}, \qquad \text{......(46)}$$

where $\frac{\partial p}{\partial n}$ is the change in pressure on the fluid for an increase in the quantity of the common component absorbed by the solid.

For solution we have

$$\frac{\partial\mu}{\partial n}dn^{\mathrm{F}}+\frac{\partial\mu^{\mathrm{F}}}{\partial p}dp=0, \qquad \text{......(47)}$$

$$\frac{\partial p}{\partial n}=-\frac{\left[\frac{\partial\mu}{\partial n}\right]^{\mathrm{F}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}} \qquad \text{......(48)}$$

giving the change in pressure on the fluid with increase in the quantity of the solid dissolved in the fluid. These expressions are not very useful in evaluating $\frac{\partial p}{\partial n}$, as $\frac{\partial\mu}{\partial n}$ is not a quantity directly measured experimentally. They are useful, however, in connections with evaluation of other derivatives.

3.4. *Miscellaneous*

In the expressions so far obtained, the pressure on the fluid has been assumed not to act on the solid. If it does act on the solid, then in the case of solution it can only be permitted to act on two opposite faces of the solid as the equilibrium pressure is different for the three opposite pairs of faces. If the normal to the faces on which it acts is denoted by m, which may take values 1, 2 or 3, then the pressure exerts tensile stress on these faces equal to $X_{mm}=-p$. The value of $\frac{\partial p}{\partial X_{kl}}$ (X_{mm} varying so as to equal $-p$) is equal to $\frac{\partial p}{\partial X_{kl}}$ (X_{mm} constant) divided by the factor $\left(1+\frac{\partial p}{\partial X_{mm}}\right)$. In the case of swelling, the pressure may be allowed to act on all three faces and the corresponding dividing factor becomes $\left(1+\sum_1^3\frac{\partial p}{\partial X_{ii}}\right)$.

The above corrections apply for $mm\neq kl$. For $mm=kl$, the expressions of (3.1)-(3.3) apply without correction, it being understood that X_{mm} is the resultant of the applied normal stress and pressure of the fluid. The same dividing factors should be used to correct $\frac{\partial p}{\partial T}$.

The derivatives of the pressure on the fluid with respect to the other variables of the system—stress on solid, temperature, and composition of phases—have now been given. These derivatives may be used to calculate other quantities. For example, the variation in composition with temperature may be computed as follows:

$$dp=\frac{\partial p}{\partial X_{kl}}dX_{kl}+\frac{\partial p}{\partial T}dT+\frac{\partial p}{\partial n}dn. \qquad \text{......(49)}$$

For constant p and X_{kl} these are

$$\frac{\partial n}{\partial T}=\frac{-\frac{\partial p}{\partial T}}{\frac{\partial p}{\partial n}}. \qquad \text{......(50)}$$

For swelling, using equations (39) and (46),

$$\frac{\partial n}{\partial T}=\frac{-\frac{1}{T}\frac{\partial L}{\partial n}}{\left[\frac{\partial \mu}{\partial n}\right]^{\mathrm{S}}}. \qquad \ldots\ldots(51)$$

For solution at low stress, using equations (40) and (48) and neglecting the first term in the numerator of (40),

$$\frac{\partial n}{\partial T}=\frac{\frac{1}{T}\frac{\partial L}{\partial n}}{\left[\frac{\partial \mu}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(52)$$

For $\frac{\partial L}{\partial n}>0$ and for $\left[\frac{\partial \mu}{\partial n}\right]^{\mathrm{S}}$ and $\left[\frac{\partial \mu}{\partial n}\right]^{\mathrm{F}}$ of the same sign, increasing the temperature reduces the amount of the common component absorbed by the solid, (i.e. reduces the amount of swelling), but increases the amount absorbed by the liquid, i.e. increases solubility.

In a similar manner, we have for variation of the composition with stress

$$\frac{\partial n}{\partial X_{kl}}=\frac{-\frac{\partial p}{\partial X_{kl}}}{\frac{\partial p}{\partial n}}, \qquad \ldots\ldots(53)$$

and for variation of temperature with stress

$$\frac{\partial T}{\partial X_{kl}}=\frac{-\frac{\partial p}{\partial X_{kl}}}{\frac{\partial p}{\partial T}}. \qquad \ldots\ldots(54)$$

If there is only one component in each phase, both $\frac{\partial V^{\mathrm{S}}}{\partial n}$ and $\frac{\partial V^{\mathrm{F}}}{\partial n}$ are specific volumes. With this interpretation, equation (54) can be used to obtain the stress coefficient of melting temperature of a pure substance.

The difference in derivatives for the case of solution when this takes place off a stressed face or a stress-free face is interesting. In the latter case, the effect of stress on the free-energy change is only manifest in the loss of the strain energy of the material transferred to the fluid phase. Strain energy is positive for positive or negative stresses, and so the effect on the independent variable, e.g. the pressure on the fluid phase, is independent of the sign of the stress. In the case of solution from a stressed face, the potential energy of the external forces is changed and the change is of opposite sign for tension and compression; hence the effect on the pressure due to this effect depends on the sign of the external forces.

It shoud be noted that all the terms in the expressions obtained are functions of the independent variables. For example, the term $\frac{\partial e_{11}}{\partial n}$ in expression (26) is a function of temperature, of concentration and of the stress. It is not sufficiently accurate to substitute for it its value at zero stress. Some idea of its stress dependence may be obtained as follows. If we divide the strain into a strain due to swelling in the absence of stress ($=e_{sw}$) and a strain due to stress in the absence of swelling ($=e_{st}$) and if we neglect interaction of swelling and stress, we may write

$$e = e_{sw} + e_{st}$$
$$= e_{sw} + \frac{X}{E},$$
$$\frac{\partial e}{\partial n} = \frac{\partial e_{sw}}{\partial n} - \frac{X}{E^2}\frac{\partial E}{\partial n}.$$

For spruce under tension along the grain and at about 15% moisture content, taking the unit for n as 1% of weight of dry solid, $\frac{\partial e_{sw}}{\partial n} \approx 0{\cdot}0001$, whereas at 20,000 lb./sq. in. stress $-\frac{X}{E^2}\frac{\partial E}{\partial n} \approx 0{\cdot}00035$, a value three and a half times $\frac{\partial e_{sw}}{\partial n}$. For a fabric-reinforced plastic, stressed to 15,000 lb./in.² in the plane of the laminations, similar figures are $\frac{\partial e_{sw}}{\partial n} = 0{\cdot}001$ and $-\frac{X}{E^2}\frac{\partial E}{\partial n} = 0{\cdot}0005$. In compression e_{st} has opposite sign to e_{sw}, and it is quite possible that $\frac{\partial e}{\partial n}$ may be negative at high compression stresses. In a similar manner $\frac{\partial e_{11}}{\partial T}$ in expression (38) will be affected by the change in Young's modulus with temperature.

The derivatives given in §§ 3.1–3.3 are obtained by exact thermodynamic analysis and include terms omitted by other authors. For example, Barkas (1945), by using the method of thermal cycles, has derived an expression for $\frac{\partial p}{\partial X_{11}}$ similar to that in equation (26) but his expression omits the first two terms of the numerator. It is therefore of interest to estimate the error due to this. The following figures are for a typical fabric-reinforced bakelite material swelling due to water absorption from a vapour phase. The swelling coefficients are estimated for a stress of 15,000 lb./sq. in. $\frac{\partial e_{11}}{\partial n} = 0{\cdot}0015$, $\frac{\partial e_{22}}{\partial n} = 0{\cdot}001$, $\frac{\partial e_{33}}{\partial n} = 0{\cdot}01$, the unit for n being 1% change in moisture content estimated on dry weight. $\frac{\partial e_{11}}{\partial X_{11}} = 6{\cdot}7 \times 10^{-7}$, $\frac{\partial e_{22}}{\partial X_{11}} = -3{\cdot}35 \times 10^{-7}$, $\frac{\partial e_{33}}{\partial X_{11}} = -1{\cdot}67 \times 10^{-7}$, the unit for X_{11} being 1 lb./sq. in. At a stress of 15,000 lb./sq. in. the sum of the first two terms of the numerator of equation (26) is about 8% of the third term, so that in this somewhat extreme case Barkas's expression is in error by some 8%. For spruce subject to tension along the grain the following figures are estimates of values at 20,000 lb./sq. in. stress, the units for n and X being as before. The moisture content is about 15%.

$$\frac{\partial e_{11}}{\partial n} \approx 0{\cdot}00045, \qquad \frac{\partial e_{22}}{\partial n} \approx 0{\cdot}003, \qquad \frac{\partial e_{33}}{\partial n} \approx 0{\cdot}0015.$$

$$\frac{\partial e_{11}}{\partial X_{11}} = 1 \times 10^{-6}, \qquad \frac{\partial e_{22}}{\partial X_{11}} = -4{\cdot}5 \times 10^{-7}, \qquad \frac{\partial e_{33}}{\partial X_{11}} = -5{\cdot}4 \times 10^{-7}.$$

At a tensile stress of 20,000 lb./sq. in. the sum of the first two terms of the numerator of equation (26) is about 20% of that of the third term. This is of course an extreme case, as the stress chosen is about equal to the breaking stress of the wood. The error is proportional to the stress.

For spruce subject to simple shear along the grain (see equation (30)) $\frac{\partial e_{12}}{\partial X_{12}}$ has a value of about 2×10^{-5} at a stress about equal to the shear strength of the material (taken as 1000 lb./sq. in.). The value of $\frac{\partial V}{\partial n}X_{12}\frac{\partial e_{12}}{\partial X_{12}}$ in equation (30) is therefore about equal to the value of $\frac{\partial V}{\partial n}X_{11}\frac{\partial e_{11}}{\partial X_{11}}$ of equation (26). If we take $\frac{\partial e_{12}}{\partial n}=-\frac{X_{12}}{G^2}\frac{\partial G}{\partial n}$, and if we assume that $\frac{1}{G}\frac{\partial G}{\partial n}$ equals $\frac{1}{E}\frac{\partial E}{\partial n}$, then the value of $\frac{\partial e_{12}}{\partial n}$ at 1000 lb./sq. in. shear stress is about equal to the value of $\frac{\partial e_{11}}{\partial n}$ at 20,000 lb./sq. in. tensile stress. The effect of shear on vapour pressure in this case is of the same order as that of tensile stress in the direction of the grain.

If spruce is subject to a tension at right angles to the grain, the following figures apply:

$$\frac{\partial e_{33}}{\partial X_{33}}\approx10^{-5},\qquad \frac{\partial e_{11}}{\partial X_{33}}\approx-0\cdot03\times10^{-5},\qquad \frac{\partial e_{22}}{\partial X_{33}}=-0\cdot55\times10^{-5}.$$

At a tensile stress of 500 lb./sq. in. (of the order of the breaking stress) the error in omitting the first two terms of the denominator of expression (26) is of the order of 1%. For this case Barkas's expression is sufficiently accurate.

The change in equilibrium moisture content with stress may be estimated using equation (53). For spruce at 18° c. and in equilibrium with 60% relative humidity and subjected to tensile stress along the grain, the proportional change in moisture content per lb./sq. in. stress $\left(\frac{1}{n}\frac{\partial n}{\partial X_{11}}\right)$ is about $0\cdot1\times10^{-6}$ at zero stress, rising to about $0\cdot5\times10^{-6}$ at 20,000 lb./sq. in. tensile stress. To calculate the latter figure, the stress-free value of $\frac{\partial p}{\partial n}$ has been used. The value of $\frac{\partial p}{\partial n}$ at 20,000 lb./sq. in. may differ somewhat from this. For the fabric-reinforced bakelite, the value of $\frac{1}{n}\frac{\partial n}{\partial X_{11}}$ is estimated to be about $0\cdot2\times10^{-6}$ at zero stress, rising to about $0\cdot3\times10^{-6}$ at 15,000 lb./sq. in. tension in the direction of the laminations.

§4. DERIVATIVES OF ENTROPY FOR SOLID-FLUID EQUILIBRIUM

The discussion is confined to equilibrium between one- and two-component phases. The expressions in 3.2 for the change in pressure on the fluid phase with temperature involve the latent heat of transfer of the common component from the solid phase to the fluid phase. It is therefore of interest to compute the derivatives of latent heat with respect to the explicit variables of the system. There is, however, a difficulty in obtaining results having much generality. When one of the explicit variables of the system is varied, another must be sumultaneously varied to ensure equilibrium. Thus it is not possible to compute true partial derivatives of latent heat. For example if, when the stress on the solid phase is varied, the pressure on the fluid phase is varied to maintain equilibrium, we have

$$\frac{\partial\left(\frac{\partial L}{\partial n}\right)}{\partial X_{kl}} = \frac{\partial}{\partial X_{kl}} T\left(\frac{\partial S^{\mathrm{F}}}{\partial n} - \frac{\partial S^{\mathrm{S}}}{\partial n}\right) \qquad \ldots\ldots(55)$$

$$= T\left[\frac{\partial}{\partial p}\frac{\partial S^{\mathrm{F}}}{\partial n}\frac{\partial p}{\partial X_{kl}} - \frac{\partial}{\partial X_{kl}}\frac{\partial S^{\mathrm{S}}}{\partial n}\right]. \qquad \ldots\ldots(56)$$

On the other hand if, to maintain equilibrium, we altered the composition of the fluid phase, we would obtain

$$\frac{\partial\left(\frac{\partial L}{\partial n}\right)}{\partial X_{kl}} = T\left[\frac{\partial}{\partial n}\frac{\partial S^{\mathrm{F}}}{\partial n}\frac{\partial n}{\partial X_{kl}} - \frac{\partial}{\partial X_{kl}}\frac{\partial S^{\mathrm{S}}}{\partial n}\right], \qquad \ldots\ldots(57)$$

or again, we could alter the temperature or another of the stresses on the solid phase. If the latter, then we obtain (calling the second stress which is varied X_{mn})

$$\frac{\partial\left(\frac{\partial L}{\partial n}\right)}{\partial X_{kl}} = T\left[-\frac{\partial}{\partial X_{kl}}\left(\frac{\partial S}{\partial n}\right)^{\mathrm{S}} - \frac{\partial}{\partial X_{mn}}\left(\frac{\partial S}{\partial n}\right)^{\mathrm{S}}\right]. \qquad \ldots\ldots(58)$$

In the circumstances, rather than give a multitude of formulae applying to particular cases, it seems best to compute the derivatives of entropy of the two phases with respect to their explicit variables separately. Derivatives of any particular type of latent heat may then be obtained by formulae of the type (56) to (58).

4.1. *Variation of entropy of solid phase with stress*

Equating the differentials of the first coefficient of equation (8) with respect to stress to that of the second coefficient with respect to temperature, we obtain

$$\frac{\partial S^{\mathrm{S}}}{\partial X_{kl}} = \left[\frac{\partial}{\partial X_{kl}}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial T}\right) - \frac{\partial}{\partial T}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial X_{kl}}\right)\right]^{\mathrm{S}}, \qquad \ldots\ldots(59)$$

$$\frac{\partial}{\partial X_{kl}}\frac{\partial S}{\partial n} = \frac{\partial}{\partial n}\frac{\partial S}{\partial X_{kl}} = \left[\frac{\partial}{\partial n}\left\{\frac{\partial}{\partial X_{kl}}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial T}\right) - \frac{\partial}{\partial T}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial X_{kl}}\right)\right\}\right]^{\mathrm{S}}. \qquad \ldots\ldots(60)$$

For simple direct stress this becomes

$$\frac{\partial}{\partial X_{11}}\frac{\partial S^{\mathrm{S}}}{\partial n} = \frac{\partial}{\partial n}\left[\frac{\partial V}{\partial X_{11}}X_{11}\frac{\partial e_{11}}{\partial T} + V\frac{\partial e_{11}}{\partial T} - \frac{\partial V}{\partial T}X_{11}\frac{\partial e_{11}}{\partial X_{11}}\right]^{\mathrm{S}}.$$

For an isotropic material having shear modulus G

$$\frac{\partial}{\partial X_{11}}\frac{\partial S^{\mathrm{S}}}{\partial n} = \frac{\partial}{\partial n}\left[V\frac{\partial e}{\partial T}\left(1 - \frac{X_{11}}{G}\right)\right]^{\mathrm{S}}. \qquad \ldots\ldots(61)$$

In the case of swelling, $\frac{\partial}{\partial n}$ means the change in the quantity inside the square brackets due to addition of unit quantity of the common component to a large quantity of the solid.

In the case of solution $\frac{\partial}{\partial n}$ may be omitted if, instead of V, the specific volume $\overline{V}$ is written.

4.2. *Variation in entropy of fluid phase with pressure*

Differentiating the first coefficient in expression (22) with respect to pressure and the second coefficient with respect to temperature, we obtain

$$\frac{\partial S}{\partial p} = -\frac{\partial V}{\partial T} \qquad \text{......(62)}$$

and

$$\frac{\partial}{\partial p}\frac{\partial S}{\partial n} = -\frac{\partial}{\partial n}\frac{\partial V}{\partial T} = -\frac{\partial}{\partial T}\frac{\partial V}{\partial n}. \qquad \text{......(63)}$$

For the swelling case in which the fluid is a pure substance, expression (63) is simply minus the temperature coefficient of the specific volume of the fluid. For the solution case, expression (63) is minus the temperature coefficient of the selling of the fluid when unit quantity of the common component is mixed with a large quantity of fluid.

4.3. *Variation of entropy with temperature for solid and fluid phases*

$\frac{\partial}{\partial n}\frac{\partial S}{\partial T} = \frac{\partial}{\partial n}\frac{C_p}{T}$, where C_p is the heat capacity at constant pressure or stress.

For a pure phase $\frac{\partial}{\partial n}(C_p)$ means the heat capacity per unit mass, that is the specific heat. For a two-component phase it requires careful definition. To measure it, it is first necessary to find the heat required to raise the temperature of a large quantity of the phase by one degree of temperature. Then let the phase absorb unit quantity of the common component. At the initial temperature and pressure again find the heat to raise the phase by one degree. $\frac{\partial}{\partial n}(C_p)$ is the difference between these two quantities of heat.

4.4. *Variation of entropy with composition*

The variation of specific entropy with composition is $\frac{\partial^2 S}{\partial n^2}$. This can only be obtained via latent-heat measurements, and no useful thermodynamic relations expressing it in terms of other parameters which can be obtained from direct experiments are possible.

§ 5. THERMODYNAMIC RELATIONS FOR INDEPENDENT VARIABLES NOT INCLUDING STRESS

So far the independent variables of the solid phase have been temperature, composition, and stress, and for the fluid phase, temperature, composition, and pressure. Instead of stress we may use strain as independent variable. Equation (5) gives the work done in terms of strain as explicit variable. Substituting this in (1) we obtain

$$dF = -SdT + V\Sigma X_{ij}de_{ij} + \mu_1 dn_1 + \mu_2 dn_2. \qquad \text{......(64)}$$

Some of the derivatives with respect to strain, similar to those with respect to stress, which are given in §§ 3 and 4, will now be given. The method is formally equivalent to that used in these sections. The independent variables are temperature, composition strain of the solid, and pressure in the fluid.

Instead of equation (24) we have

$$\frac{\partial p}{\partial e_{kl}} = \frac{\left[\frac{\partial}{\partial n}(V\Sigma X_{ij})\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(65)$$

For the cases of swelling and solution, $\partial V/\partial n$ has the significance already discussed. Instead of equation (37) we have

$$\frac{\partial p}{\partial T} = \frac{\frac{\partial}{\partial n}\left(\frac{L}{T}\right)}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(66)$$

Instead of equation (60) we have

$$\frac{\partial}{\partial n}\frac{\partial S}{\partial e_{kl}} = -\frac{\partial}{\partial n}\left[\frac{\partial}{\partial T}(V\Sigma X_{ij})\right]^{\mathrm{S}}. \qquad \ldots\ldots\ldots(67)$$

Instead of stress and strain we might also use force and displacement as independent variables. By comparing expressions (2) and (5) and (4) and (6) it can be seen that derivatives with respect to force and displacement can be obtained from those with respect to stress and strain by writing 1 for V^{S}, P_{ij} for X_{ij} and x_{ij} for e_{ij}; for example the expression equivalent to (24) is

$$\frac{\partial p}{\partial P_{kl}} = \frac{\left[\frac{\partial}{\partial n}\left(\Sigma P_{ij}\frac{\partial x_{ij}}{\partial P_{kl}}\right) - \frac{\partial}{\partial P_{kl}}\left(\Sigma P_{ij}\frac{\partial x_{ij}}{\partial n}\right)\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}} \qquad \ldots\ldots(68)$$

that equivalent to (26) is ;

$$\frac{\partial p}{\partial P_{11}} = -\frac{\left[\frac{\partial x_{11}}{\partial n}\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(69)$$

REFERENCES

Barkas, W. W., 1945. *Forest Products Research Special Report* No. 6.
Gibbs, J. W., 1876. *Collected Papers* (Longmans & Co.).
Warburton, F. W., 1946. *Proc. Phys. Soc.*, **58**, 585.

THE KEW RADIO SONDE

By E. G. DYMOND,
University of Edinburgh

MS. received 23 *December* 1946

ABSTRACT. The British radio sonde is a system for telemetering indications of pressure, temperature and humidity from a free balloon to the ground. It is used on a large scale for routine observations of the upper air for meteorological forecasting.

It works on the principle of a varying inductance changing the note of an audio-frequency oscillator, which modulates the radio transmitter. The design of airborne instrument, ground receiving apparatus and calibrating plant is described. An account is given of the performance of the radio sonde, and of the errors to which it is subject in actual operation. The probable errors are in the neighbourhood of ±5 mb. and ±0°·4 c. for pressure and temperature over the atmospheric range up to 22 km. height, and ±10% relative humidity down to temperatures of −20° c., below which the hygrometer element becomes unreliable or inoperative. The reliability is high, over 95% of the soundings being successful.

§ 1. INTRODUCTION

A RADIO SONDE is a meteorological instrument which can be attached to a free balloon in order to measure pressure, temperature, and humidity during ascent. The indications are transmitted to the ground by a radio link. A network of stations using such instruments enables a three-dimensional picture of the atmosphere to be obtained, thus providing the forecaster with far more information than can be derived from surface measurements alone.

The following is an account of the radio sonde developed for the Meteorological Office during the war. It is generally known as the Kew Mark IA radio sonde. It has been briefly described, in a general review of recent meteorological developments, by Johnson (1946).

The requirements for such an instrument are that it shall measure pressure and temperature to an accuracy of at least 1% of the range, humidity to between 5 and 10%, that it shall be sufficiently light to be carried by a balloon to a height of 16 km. on a majority of occasions, and that its cost shall be sufficiently low to allow of large-scale use, even when the chance of recovery after a flight is small.

The progenitor of the Kew radio sonde was an instrument designed by Thomas (1938) at the National Physical Laboratory. This gave continuous readings of pressure and temperature but there was no means of measuring humidity. It incorporated two audio-frequency oscillators with variable inductors, each of which was controlled by a meteorological element in such a way that the frequency of oscillation was a function of pressure or of temperature. The two oscillators simultaneously modulated a radio transmitter. Reception on the ground was by a normal communications receiver, the output from which was matched in frequency by ear with that of a calibrated variable oscillator. The two audio-frequencies were sufficiently spaced so that no confusion arose between them. Power for the balloon-borne instrument was provided by dry cells.

This radio sonde was not satisfactory for two reasons. No measurement of humidity was possible, nor was the frequency stability adequate to give acceptable accuracy. A modified form was produced by Thomas in 1939 in which the inductors were redesigned and a humidity unit added. The most important change was the substitution of mumetal for silicon iron as the material of the inductor cores. The frequency stability was much increased and adequate accuracy in the air could be expected. But the instrument still suffered from the following disadvantages:—

1. It was very heavy, weighing 2920 grams, of which the battery represented 960 grams.
2. The temperature unit, though possessing the desirable feature of very quick response to changes in temperature, was insufficiently stable. In particular it was unduly sensitive to gravity and to changes in tilt. As a radio sonde

is apt to swing through quite large arcs while ascending, unpleasantly large errors would be introduced.

3. The humidity unit was unsuited to routine use. It operated by causing a change in the pressure calibration in steps as the humidity varied.

For the above reasons the Thomas instrument was not adopted for use in the Meteorological Office, but a new design was called for, operating on the same general principle of variable inductance, but avoiding the undesirable features described above.

General features of design

As the humidity unit of Thomas was quite impracticable when applied to an instrument required for large-scale use, owing to the large number of separate pressure calibrations required, it was decided to add a third inductor, similar to those for pressure and temperature. But if a third audio-frequency circuit were added, the weight and complexity of the instrument would be increased to an inadmissible extent. Furthermore, reception would become difficult, as three audio-frequencies would be transmitted instead of two. Trials with an experimental model incorporating three oscillators confirmed this view and showed that this line of attack was not feasible.

Accordingly, in the Kew instrument, there is only one audio oscillator. Each inductor is in turn connected into the circuit by means of a switch which is driven by a wind vane. The vane rotates in the vertical slip stream. Two good features of the Thomas design were sacrificed, continuous recording and total lack of moving parts other than those of the meteorological elements themselves; but it now became possible to measure humidity in a simple manner and also to reduce the number of valves in the circuit. As an indication of the extent of simplification, the total weight was reduced from the 2920 grams of the Thomas model to 1400 grams. This, however, was partly achieved by the adoption of another type of battery.

The individual meteorological elements, with their inductors, form a unit which is detachable, as in the Thomas instrument. This is of importance in calibration, as the test chambers required are very much smaller than if the whole radio sonde had to be placed within them. In order to cheapen production the three units were made as nearly as possible alike.

§ 2. DESCRIPTION

A schematic diagram of the radio sonde is given in figure 1, which shows the three variable inductors operated respectively by an aneroid capsule for pressure, a bimetallic strip for temperature and a strip of goldbeater's skin for humidity. A photograph of the complete instrument without its container or battery is shown in figure 2. The battery rests on the lower circular panel. A cylindrical case of bakelized cardboard protects the two panels. The three inductors, which project from the case into the air stream, are shown in their shields to protect them from solar radiation. In figure 3 is given the circuit diagram. All three valves are of type HL 23, taking 55 ma. filament current.

The audio oscillator

This consists of a Hartley oscillator (V_1) with a frequency range of 700 to 1000 cycles per sec. C_1 is the condenser, of 0·07 mF. capacity, which with the inductor in circuit at the moment forms the tank circuit of the oscillator. The precision of the instrument depends largely on the stability of this condenser. As an overall constancy of 0·5 c./s. in frequency is aimed at, this condenser must maintain its capacity to well within one part per thousand even at the lowest temperatures likely to be reached. Silvered mica condensers were first used, but difficulties in production of the necessary quantities led later to the adoption of a clamped mica type* in which a very low temperature coefficient of capacity could be achieved. The average value of this coefficient in production samples is 29×10^{-6} per °c., though figures of 10×10^{-6} are possible in selected condensers. Many types of silvered mica condenser have temperature coefficients up to 50×10^{-6} per °c.

It is obviously important that the frequency should change as little as possible with battery voltage. A large measure of stabilization has been achieved by the method adopted by Thomas, in which a combination C_2R_2 is inserted between the oscillating circuit and the driving valve. The large impedance of this combination tends to swamp small changes in valve constants which might cause frequency variation. Frequency instability due to changes in grid current is discussed in a later section. Values of C_2 and R_2 are chosen empirically to give the best performance. It is possible to reduce the variations due to changes in anode voltage to zero over a small range, but those due to low-tension changes cannot be compensated at the same time. A compromise is arrived at in which simultaneous alteration of high and low tension supplies causes frequency variations of opposite sign. Typical figures are −0·12 c./s. per volt for high tension and +3·5 c./s. per volt for low tension changes. In actual practice this means an overall variation of +0·3 c./s., due to the average drop in battery voltage to be expected during a flight.

An important component of the circuit is the condenser C_9. This serves to decouple the grid of V_1 from radio-frequency oscillations arising in the transmitter. Without this condenser, it is found that there is some feed-back through the modulator stage. The radio-frequency voltage appearing on the grid of V_1 is rectified and biases it back sufficiently to alter the audio frequency. This effect is particularly disturbing as it depends in magnitude on which inductor is in

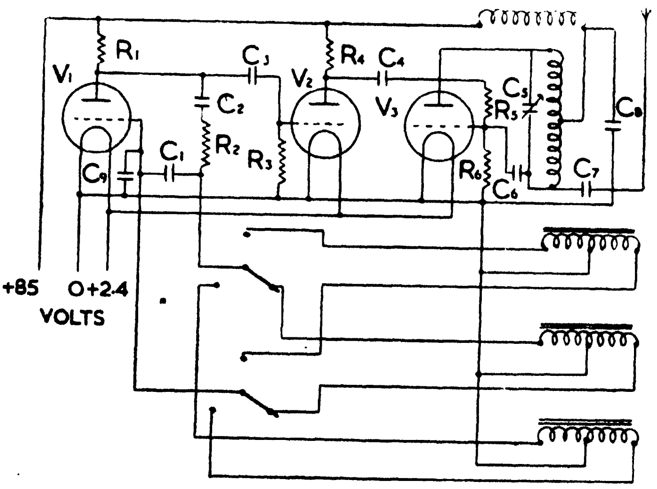


Figure 3. Circuit diagram.

$R_1 = R_2 = R_5 = R_6 = 22$ K.
$R_3 = R_4 = 47$ K.
$C_1 = 0{\cdot}07$ mF.
$C_2 = 0{\cdot}01$ mF.
$C_3 = C_4 = 0{\cdot}001$ mF.
$C_5 = 5 - 20$ pF.
$C_6 = 20$ pF.
$C_7 = 5$ pF.
$C_8 = 100$ pF.
$C_9 = 500$ pF.

---

* Designed by the Dubilier Condenser Co, Ltd.

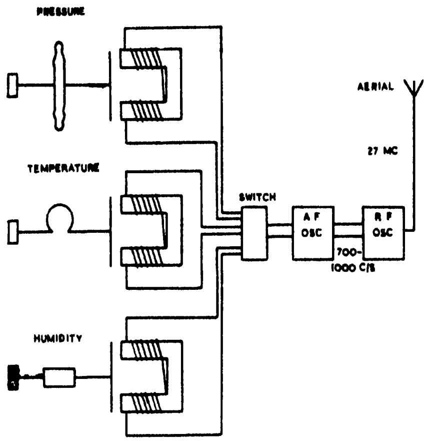


Figure 1. Schematic diagram of the Kew radio sonde

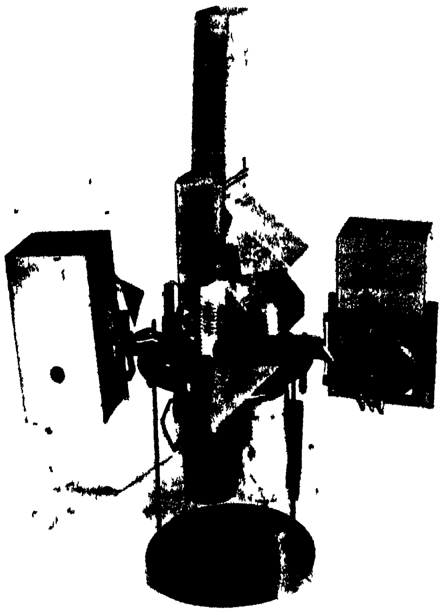
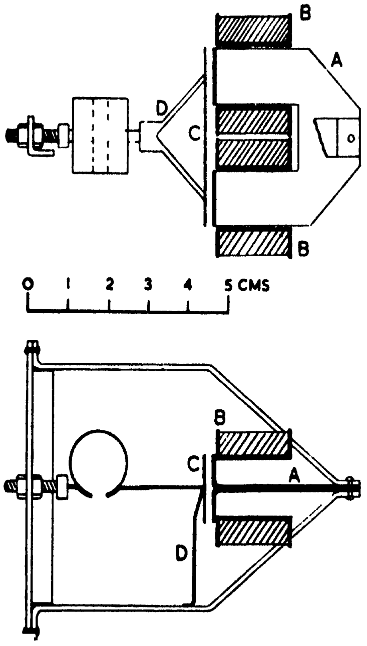


Figure 2. The radio sonde, without container or battery.



. An inductor, with temperature element. Plan and elevation. Slightly diagrammatic.

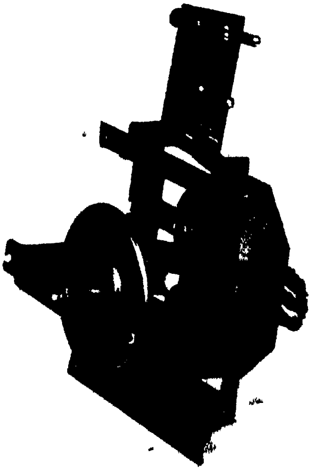


Figure 5. The pressure unit, without radiation shielding.

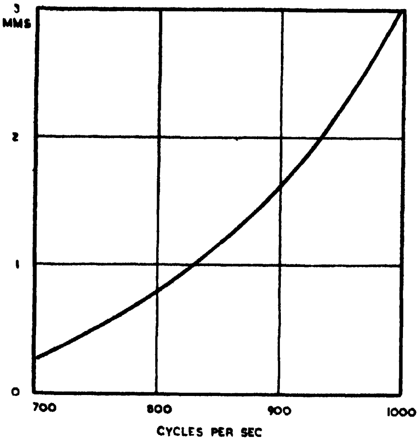


Figure 6 Relation between width of gap in the magnetic circuit and oscillator frequency

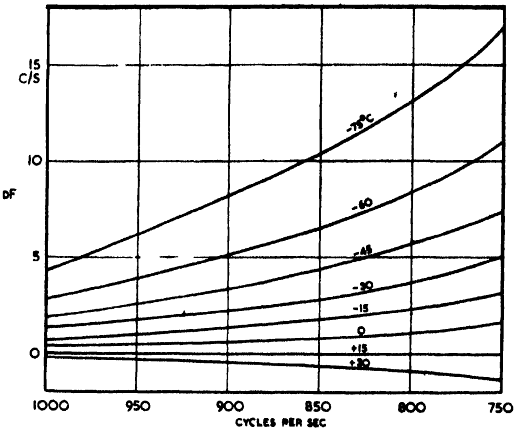


Figure 7 Frequency change, *dF*, against oscillator frequency, for various temperatures

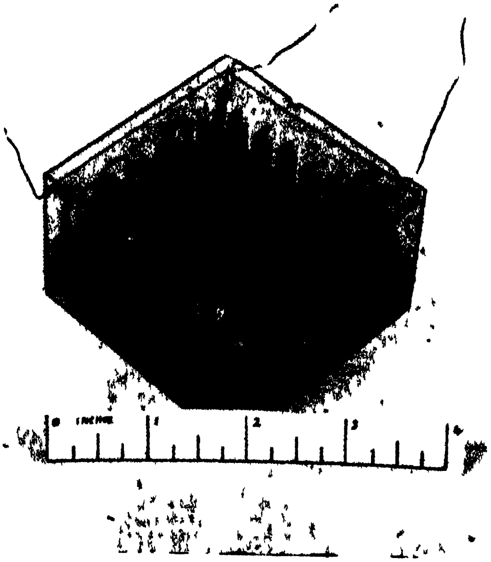


Figure 8 The battery, providing 86 volts high tension and 2 4 volts low tension.

circuit, due to differences in the r.f. impedance of the leads, and on the loading of the r.f. stage. A decoupling condenser of 500 pF. capacity is sufficient to make all variations due to r.f. interaction less than 0·1 c./s.

During calibration, only the inductor units are placed in the cold chamber. It is therefore necessary that the remainder of the oscillator circuit as a whole should be insensitive to temperature, as during an actual ascent it will be subjected to cold conditions. This question has been studied by reversing the normal calibrating conditions. The body of the instrument is placed within the calibrating chamber, with the inductor outside at room temperature, connections being made with leads through the chamber wall. It is found that a temperature change from +15 to −60° C. causes a frequency shift of +0·4 c./s. As the condenser C_1 alone should give about +1·0 c./s. in these conditions, the rest of the circuit, including the valve, produces a shift of about −0·6 c./s. The net effect is sufficiently small to be neglected.

Modulator stage

The valve V_2 is interposed between audio- and radio-frequency oscillators to impose a constant load on the former and so maintain its stability. Some measure of amplification also occurs in this stage.

Radio transmitter

This works in the frequency range 26 to 30 Mc./s. The single valve oscillator V_3 is grid-modulated by the output of V_2. The depth of modulation can be controlled by the magnitudes of the coupling resistor and condenser, R_5 and C_4, between V_2 and V_3. It is found that two states of modulation are possible. In one, the modulation depth is 30% or less, with little distortion of the audio-frequency wave-form. In the other, the oscillator V_3 is over-modulated, as during part of the cycle it is completely cut off. Depths between 30 and 100% cannot be obtained. The state of over-modulation was chosen to obtain the maximum signal strength. The wave-form is much distorted, but this is of no consequence.

With this type of circuit there is a large measure of frequency modulation in addition to the amplitude modulation. This has not been accurately measured, but the frequency swing is of the order of +50 Kc./s. When receivers with narrow pass bands are used, there is little gain in signal strength when the modulation depth is increased from 30 to 100%. But in addition to its use for measuring pressure, temperature and humidity, the radio sonde is also used for measuring winds, as described by Johnson. The direction-finding receivers used for this work are not as selective as those for radio sounding and the mode of greater modulation gives some advantage in signal strength.

The aerial is end-fed, half a wave-length long, and is attached alongside the suspending cords from the balloon. The aerial current as measured by a thermal milliammeter inserted in the midpoint of the aerial is of the order of 15 to 25 ma. It depends largely on the particular valve, as these vary widely in the efficiency of oscillation. In practice, the radio-frequency valve is selected to give a minimum output of 15 ma., with an anode voltage of 85 volts. Assuming an aerial impedance of 70 ohms, these currents correspond to a radiated power of 16 to 44 milliwatts. This power is amply sufficient. The maximum altitude of ascent is reached in

45 to 60 minutes, and it is only on very rare occasions that the instrument is carried by the wind more than 100 miles (160 km.) in this time. But instances have been recorded of the radio sonde from one station being heard by another when over 175 miles (280 km.) distant. It appears that limitation in range is due rather to passing of the transmitter below the horizon than to fading of the signal owing to distance alone.

As the radio transmitter is subjected to a large change in temperature during flight, its frequency suffers a steady drift. This is minimized by use of an air-dielectric condenser C_5 as the tuning condenser. Drifts are also caused by the voltage drop of the battery. In practice the frequency shift during a flight averages 100 and rarely exceeds 200 Kc./s.

Meteorological units

These, as already mentioned, are separate and can be plugged into the main body of the instrument. Whereas the battery and circuit elements are enclosed for protection against the weather and for thermal insulation, the meteorological elements and their inductors are fully exposed to the air. This is of importance as the inductors are somewhat sensitive to temperature, and in order to apply corrections accurately their temperature must be known with some precision. The inductors are all alike, with the exception of that in the humidity unit, in which the number of turns of the coils is reduced. This is for the purpose of slightly raising its frequency range, so as to separate it to some extent from those of the other two units.

It is important in the design of these units that they shall be mechanically robust, so that they do not distort under the shocks of transport, and that there should be a minimum of metal in regions where stray magnetic flux is strong. This flux sets up in all the metal parts eddy currents, which react very unfavourably on the frequency stability.

An inductor is shown diagrammatically in plan and elevation in figure 4. A photograph of the pressure unit without radiation shield is given in figure 5. The core A is made of six mumetal stampings, each 0·005 in. (0·127 mm.) thick, whose ends are turned up to form flat pole pieces. The coils B are wound on moulded formers, each with 1200 turns of No. 38 S.W.G. copper wire, insulated with fused cellulose acetate. In the case of the humidity unit, the coils are wound with 1100 turns only. The coils are placed as close to the pole pieces as possible, to reduce the leakage flux. The resistance of the two coils is about 120 ohms at 15° c., but falls to 80 ohms at $-60°$ c.

The moving armature C is a single stamping of mumetal 0·005 in. (0·127 mm.) thick, supported by a nickel silver stamping D of the same thickness. One portion of this stamping serves as a spring hinge for the armature; to the other is attached the meteorological element. The flux density in the armature is much greater than in the core, and it is necessary that it should be of the highest permeability. This is required not so much to obtain a high value of inductance as to reduce the losses in the mumetal. These losses, partly due to hysteresis and partly to eddy currents, are smaller the higher the permeability. It is therefore important that the strip from which the armature is cut should have been rolled in the direction of its long axis, in order to obtain the best magnetic qualities, and that after heat treatment the armature should be treated carefully. Mumetal in its high permeability state is very sensitive to mechanical handling. It is found that the stamping D may be soldered on without affecting the permeability, but that spot welding is definitely deleterious.

It will be noted from figure 3 that grid current flows through one coil of the inductor. This current, though small, produces a permanent flux in the magnetic circuit which reduces the incremental permeability. Any change in its value will therefore alter the oscillation frequency. An attempt was made to improve the frequency stability by using a grid leak with condenser coupling, thus removing the grid current from the coil. It was found, however, that the stability was worsened. Apparently when the supply voltages are altered the effect of changing grid current partly compensates for the variations in the other characteristics of the valve.

A typical curve showing the variation of frequency with air gap in the magnetic circuit is given in figure 6. The useful range of movement of the armature is about 2·5 mm. The sensitivity varies from 40 c./s. per mm. at 1000 c./s. (wide gap) to 230 c./s. per mm. at 700 c./s. But the precision of measurement cannot be expected to increase in the same proportion, as at the low-frequency end the gap is so small that slight distortions in the frame of the unit will have proportionately a bigger effect. The design is such that meteorological conditions on the ground correspond to high frequencies, both for pressure and temperature.

The pressure unit

The sensitive element is an aneroid capsule of steel, K monel or beryllium copper. A very low value of elastic hysteresis is required, and also a nearly linear relation between deflection and pressure. The Meteorological Office specification calls for a maximum width of the hysteresis loop of 2 millibars in a complete pressure cycle of 1000 mb. amplitude.

A deflection linear with pressure, when combined with a frequency-deflection relation as shown in figure 6, gives a very desirable characteristic to the instrument. Due to the logarithmic relation between atmospheric pressure and height, a much higher pressure sensitivity is required at low than at high pressures. The Kew radio sonde does not achieve the ideal of linear height sensitivity, but approaches it more nearly than do most such instruments.

The pressure unit is sensitive to temperature and, as it is subjected to the full range of atmospheric temperature, it is necessary to evaluate corrections for this. The change dF in frequency due to temperature may be expressed as

$$dF = dF_1 + dF_2 + dF_3,$$

where dF_1 is due to the aneroid capsule itself, dF_2 to the inductor coils, and dF_3 to the mumetal in armature and core. dF_1 is caused by the temperature coefficient of the elastic constant of the capsule. The effect of this is greatest at ground level, when the stress on the capsule is a maximum, and it becomes zero at the top of the atmosphere. The temperature changes at ground level are, however, comparatively small, and so the importance of dF_1 increases with height to a maximum at about 200 mb. pressure and decreases again with further increase in height. It is possible to introduce some form of compensation in the capsule, but this course has not been followed as it is found that the contribution of dF_1 to the whole effect is small.

The resistance of the coils changes by about 30% between +15 and −60° c. This alteration in the resistance of the oscillating circuit changes its frequency by dF_2. dF_3 is caused by the variation in hysteresis and eddy currents in the mumetal and, in a minor degree, to the change in permeability.

The total variation, dF, is not a linear function of either temperature or frequency, and it also varies widely from instrument to instrument. The average values are shown in figure 7, where dF is plotted against observed frequency for various temperatures. It is seen that in the upper atmosphere, corresponding to the region below 800 c./s., the corrections which must be applied are quite large and may reach, when converted into pressure, 20 mb. This is an important fraction of the total pressure.

Some insight into the relative magnitudes of dF_1, dF_2, and dF_3 may be gained from the following considerations :—

(*a*) By fitting an extension to hold the aneroid capsule at some distance from the rest of the inductor, the change in frequency can be observed when (1) the whole unit is cooled and (2) when only the capsule is cooled by immersion in a suitable bath. The differences give the effect of the capsule alone. Changes due to thermal contraction of the frame can be shown to be small. In a typical example it is found that $dF_1 = 2{\cdot}0$ c./s. at surface pressure for 80° c. change. At 200 mb. pressure (800 c./s.) $dF_1 = -0{\cdot}04\ dF$, and is, therefore, negligible.

(*b*) The change in resistance of the coils on cooling is measured. Resistance is now added to the oscillating circuit when cold to restore the original value. The alteration in frequency due to the restoration of resistance is measured, and this must equal dF_2. For an 80° c. change dF_2 is found to be $-4{\cdot}8$ c./s.

(*c*) From the foregoing results dF_3 is seen to be $1{\cdot}52\ dF$ at 200 mb. It is, therefore, by far the most important factor, and its variation from instrument to instrument contributes to the wide range of dF found in practice.

The effect of the mumetal is itself complex, as both core and armature have to be taken into account. The relative effect of these components can be estimated by removing the armature altogether. But this can only give rough results, as, without the armature, the flux in the core differs widely from the working conditions.

More reliable estimates can be made from the following experiment. The whole unit is cooled to a low temperature and is then suddenly placed in an air stream of about 5 metres/sec. at room temperature, while the change in frequency with time is measured, as the unit warms up. The curve connecting dF with time consists of two exponentials, with half-value periods of 5 and 25 seconds and relative amplitudes of 3 to 1 respectively. The curve of quick decay represents the change due to the armature, which is very thin and exposed to the full ventilation. The slower curve corresponds to the coils and core, which are to be expected by reason of their construction to change their temperature together.

This experiment is of importance also in assessing the accuracy of pressure measurement. The proper application of the correction dF to observed values of frequency assumes that all parts of the pressure unit are at the same temperature or, more exactly, that they have the same relative temperature distribution during flight as during calibration. It is seen from the above discussion that the greatest contribution to dF arises from the armature, which follows the air temperature very rapidly. The coils and core have a considerable lag but do not have a large effect on dF. This lag does, however, limit the ultimate accuracy with which pressure can be measured, which, as will be shown later, is of the order of 5 mb.

The temperature unit

The sensitive portion is a bimetallic strip, 0·025 mm. thick, rolled into a cylinder 1 cm. diameter by 1·6 cm. high. The bending into a circular arc sets up stresses which are only slowly relieved, in spite of repeated cycling of the element over the complete working range of +30 to −70 °c. Investigation has shown that the rate of recovery with time after bending is exponential. In the first instruments, a bimetal of brass and invar steel was used, in which the amount of creep after rolling equalled the deflection due to 3°·25 c. change; 90% of the deflection was reached in 135 days. It is not possible to anneal this material effectively, as high temperatures lead to softening of the brass and destruction of the elastic properties. A combination of ordinary and invar steel was finally adopted, which gives a creep after rolling equivalent to 1°·3 c., with a 90% period of 110 days. Although this type of bimetal can be annealed, this is not possible in the actual elements used. These are arranged to curl up with increase of temperature, and at the annealing temperature the edges of the split cylinder would meet and set up fresh stresses. In spite of having only three-quarters of the sensitivity of the brass-invar combination, the new material has been found to give markedly improved performance, due to its better elastic properties. As calibration takes place at least one and usually several months after rolling, the effect of creep can be kept quite small.

At low temperatures, the sensitivity of bimetal is reduced, owing to the increase in Young's modulus. This is countered by the increasing sensitivity of the inductor for small gaps, as shown in figure 6. The combination gives a variation of frequency with temperature which may be made linear or slightly increasing with falling temperature, depending on the sensitivity of the particular bimetal.

An important characteristic of a temperature element for radio sonde work is its speed of response. In this instrument the time for a 50% response to an instantaneous temperature change is 4·5 sec. in an air stream of 5 m./s. and of normal density. The thickness of metal forming the bimetallic strip is the controlling factor in determining the lag. It is not feasible to reduce this further without impairing stability, as the cylinder would become too weak. It is also important that the strip should be ventilated as freely as possible. The lag gives rise to a systematic error near the ground of 0°·2 c. when the instrument is rising at its normal rate in an atmosphere with a lapse rate of 6° c. per km. This error varies inversely as the square root of the air density, and will amount to 0°·45 c. at 200 mb., about the level of the tropopause. In the stratosphere the error becomes negligible, since in this region the temperature itself is practically constant.

No attempt is made to apply a correction for the effect of lag. Its variation with height leads to a very small but significant error in the lapse rate; however, since all instruments of the same type are equally affected, horizontal temperature gradients are not appreciably changed.

The temperature coefficient of the inductor itself is the same as that in the pressure unit. Errors due to the inductor, however, will only arise if the temperature distribution of the various parts is different in actual flight from that occuring during calibration. The performance of the instrument does not suggest that any perceptible error arises from this cause. As previously shown, the armature, which makes the largest contribution to dF, has a lag coefficient which is so close to that of the bimetal that the two may always be considered to be at the same temperature.

The most difficult problem in designing a temperature-measuring system for radio sondes is the prevention of radiation errors, particularly in the higher levels of the atmosphere, where solar radiation is most powerful and ventilation least effective. Not only direct radiation from the sun but also that from clouds beneath the instrument must be considered. At night there is also the possibility that the bimetal may be cooled by radiation into space.

On all these counts it is of the first importance that the surface of the bimetal shall be as highly reflecting as possible. Unfortunately, it has been found that a coating of nickel or chromium adequate to take a high polish is so thick that not only is the sensitivity reduced but the stability of the element is also impaired. But one commercial grade of bimetal* is formed of stainless steel which will itself take a high and permanent polish.

The perfect radiation shield for radio sondes has yet to be designed. There is no difficulty in protecting the sensitive element from radiation arriving at a low angle to the horizon, but the problem of dealing with high solar elevation has not been completely solved. The following requirements, some of which are mutually antagonistic, should be met:—

1. The shield must not allow solar radiation to strike the element directly.
2. It must prevent radiation reaching the element after multiple reflection within the shield.
3. The shield must not itself absorb sufficient radiation to warm appreciably the air flowing through it.
4. There must be no interference with the free flow of air past the element.

Condition 1 suggests a tall shield with narrow opening at the top, but 3 requires the reverse; 2 demands a complicated structure which will conflict with 4. The screening of the Kew radio sonde is necessarily a compromise. The present form was reached after many modifications during the last three years, and is by no means ideal, as the necessity of maintaining production continuously did not allow of drastic alterations to preceding designs. It consists essentially of a double aluminium shield which is extended in the upward direction by a thin rectangular tube (see figure 2).

It is believed that this system is fully effective in temperate regions up to the highest levels. The evidence for this statement will be discussed below, under the heading of "Performance". But in tropical regions, in the period around noon, it cannot be expected that this or any other radio sonde will give temperature readings which are not falsified by radiation effects.

The humidity unit

The sensitive element is a strip of goldbeater's skin. This material has several advantages for hygrometric measurements over the conventional hair. It is much more sensitive, it gives more reproducible readings, and its lag in conditions of changing humidity is much less. The sensitivity varies from 4 to 8% change in length, depending on the particular sample, for 100% change in relative humidity, and is independent of temperature.

As the distribution in the atmosphere of humidity with height is much more irregular than that of temperature, it is important to have an instrument with a minimum of lag. Unfortunately, the speed of response falls rapidly with falling temperature, as table 1, due to Glückauf (1947), shows.

Table 1

Temperature (° c.)	+18	0	−30	−69
Time constant (sec.)	2·4	6	60	1800

---

* Hiflex, supplied by Henry Wiggin and Co.

These results were obtained in an air stream of 5 m./s., about the speed of ascent of a radio sonde. The corresponding figure for hair at +18° c. is about 30 sec. Glückauf has also shown that the maximum speed of response at a given temperature occurs for changes in the neighbourhood of 50% relative humidity, and falls off both in very dry and in very damp air. In particular, the lag approaches infinity at 100% R.H. The lag at low temperatures limits the region of reasonable accuracy to above −20° c., and at −40° c. the material becomes useless for hygrometry.

Glückauf has also shown that goldbeater's skin exhibits a hysteresis effect when subjected to a cycle of humidity changes, which includes very dry conditions, but that it recovers its original calibration when it returns to above 70% R.H. There is no hysteresis between 70 and 100% R.H.

Thus it is seen that an instrument using goldbeater's skin leaves much to be desired. No other material, however, is available with better qualities, nor do the electric surface-resistance types first developed by Dunmore (1939) show any better performance at low temperatures.

The successful use of the material depends on the observance of the following practical points:

1. The skin must be single-ply and unvarnished.
2. The maximum working tension must be limited to 50 grams per cm. width.
3. After mounting on the unit, the skin must be seasoned for several hours in a saturated atmosphere, while subjected to its working tension. The material acquires a permanent strain under this treatment, without which it is impossible to obtain reproducible results.
4. In order to minimize the hysteresis, it is advisable to condition the element by placing it in a saturated atmosphere for 20 minutes both before calibration and use, and to calibrate from damp to dry conditions. This simulates the usual direction of humidity change during an ascent.
5. While after the conditioning process no further permanent change in length occurs in a saturated atmosphere, this is not true if the material is placed in liquid water. Therefore the strip of goldbeater's skin must be protected from rain. It has been found that passage through cloud does not affect the calibration, but prolonged exposure to extremely wet fog while preparing for an ascent has on occasions given rise to further stretching.

The inductor of the humidity unit is similar to those for the pressure and temperature, except that the coils are wound with 1100 instead of 1200 turns of wire. Owing to the limitation in accuracy imposed by the nature of the material, great precision of reading is not required and the range of frequency is limited to about 100 c./s., in the upper portion of the band. In the higher atmosphere, where humidity readings are useless, the frequency of the unit is then well separated from those of the pressure and temperature units.

The effect of temperature on the calibration can be neglected. Glückauf has shown that the calibration of the skin itself is independent of temperature, and corrections due to the inductor are unimportant above −20° c., where useful readings may be obtained. It is thus unnecessary to calibrate the unit except at room temperature, an important practical advantage.

The battery

The power supply for the radio sonde must have the following characteristics:—

1. Small weight.
2. Constant potential during discharge.
3. Long shelf life before use.
4. Relative insensitivity to low temperatures.

The form of battery most nearly conforming to these conditions is that used by Väisäla (1937). A similar design, of larger capacity, has been adopted for the Kew radio sonde. A photograph is shown in figure 8. Both high- and low-tension cells are constructed with lead peroxide positive and amalgamated zinc negative plates. The electrolyte is sulphuric acid, of density 1·27. After prolonged storage, the mercury on the zinc electrode diffuses into the body of the metal; to counter this effect about 1% of mercuric sulphate is added to the electrolyte. This provides a freshly amalgamated surface at the moment of use. The case is moulded in cellulose acetate.

Such a cell gives an e.m.f. of between 2·4 and 2·5 volts. The characteristics of the complete battery are shown in table 2.

Table 2

Type	Cells no.	Volts	Discharge max. (ma.)	Discharge working (ma.)	Capacity (ma. hrs.)	Cap./weight (milliwatt hrs./gm.)
HT Mk. I	36	86·0	10	6	12	7
LT. Mk. I	1	2·4	250	175	300	11
HT Mk. III	40	98·0	30	30	45	17
LT Mk. III	3	7·2	600	600	900	29

The Mark I battery is that used for the radio sonde. A single moulding contains both high- and low-tension cells. The weight complete with acid is 300 grams. This can be considerably reduced, as later developments have shown. As an example of what is now possible, figures are also given for the Mark III battery, developed for another type of instrument. The increased performance has been attained by increasing the amount of active paste relative to inactive grid in the positive plate, and by reduction of its thickness to the minimum required for mechanical strength. As the discharge rate is very high, the capacity is dependent on the surface and not on the volume of the plate.

At the working discharge rates given in the table, the e.m.f. remains constant to 5% for at least 1½ hours, which is usually sufficient not only for the ascent but for a large part of the descent also. This constancy is of assistance in maintaining the frequency stability of the oscillators.

The positive plate is given a special forming charge during manufacture, and is carefully washed and dried before assembly. The capacity falls with time, but the stated performance can still be obtained after 12 months' storage. There is evidence that most of the reduction takes place in the first three months; but if the battery is kept in a dry atmosphere, sealed from the air, there is no diminution in capacity, and in this condition the shelf life is indefinitely long.

The sensitivity to low temperature depends on the rate of discharge. The e.m.f. changes little, but the internal resistance rises, with fall of temperature. At low rates of discharge the cells can be used down to about −30° c., but at the maximum rate failure occurs at about −15° c. This has been shown by Marth (1944) to be due to precipitation of zinc sulphate from solution, leading to the formation of a high-resistance layer at the negative electrode. In actual practice the battery is well lagged in cellulose wadding and is placed within the case of the radio sonde. Its temperature during an ascent can be studied by laboratory experiments in which the thermal lag of the battery is measured in conditions closely approximating to those during a flight. These lead to the conclusion that if the maximum altitude is reached in 45 minutes, with an air temperature of −60° c., the battery will fall to −15° c. As radio-sonde failure due to battery trouble is rare, it is believed that these experiments give a pessimistic view. In many cases the instrument can be followed, after the balloon has burst, until the descent is complete, with no indication of battery failure.

The switch

This connects in turn each meteorological unit to the audio-frequency circuit. The contacts, which are of gold-plated phosphor bronze, are operated by a cam driven through worm and gear wheel by a three-armed windmill, similar to a cup anemometer. This rotates in the air stream created by the ascent of the instrument. Near the ground, the switch makes a complete cycle of operations in about 20 seconds, giving 6 sec. to record each meteorological element. In the stratosphere, the rate decreases to about 1 cycle per minute. As the amount of power available at high altitudes is very small, the switch and gear must operate with a minimum of friction, and no lubricant can be used owing to the low temperature. The contacts are protected by a closely fitting cover to prevent condensation of moisture given off by the battery.

Ground-receiving apparatus

The apparatus on the ground for receiving and analysing the signals of the radio sonde consists of

(*a*) Radio receiver.
(*b*) Calibrated variable audio-frequency oscillator.
(*c*) Cathode-ray oscillograph.

The audio-frequency signal, derived from the receiver, is applied to one of the pairs of plates of the oscillograph, and the output from the variable oscillator to the other pair. A stationary loop is seen on the screen when the two frequencies are equal. The oscillator can be set by this means rapidly to within 0·1 cycles/sec. of the frequency of the radio sonde, the value of which can be read off a dial.

The superheterodyne receiver and oscillograph are normal commercial products and call for no comment. The oscillator must be accurate to within 0·2 c./s. over the range 700 to 1000 c./s., a degree of precision which is not attained by any commercial instrument. A beat-frequency oscillator was first developed, but because of drifts in frequency it required standardizing against an electrically maintained tuning fork at short intervals. Ultimately a resistance-capacity oscillator * was used which had a reading accuracy of 0·2 c./s., and which was constant to this amount over periods of at least two hours. It was therefore only necessary to check the calibration against the tuning fork once before each ascent. The tuning fork itself is of Elinvar, but is not maintained at a constant temperature, so that variations of frequency due to extremes of temperature may be of the order of 0·1 c./s.

The procedure of taking observations and computing the results is as follows. One man observes and measures the frequency of each signal, corresponding to pressure, temperature, or humidity, as they occur in turn. He plots against time

---

* Developed by Muirhead and Co. Ltd.

each observed frequency, a special clock graduated in 1/20 minutes being used for timing. The scales of the plotting chart are so adjusted that with readings taken to 0·5 c./s. no interpolation is required. This is necessary as the time for observing and plotting each point is only 6 sec. The record therefore consists of a series of dots, running in three lines representing frequencies of the pressure, temperature and humidity units. There is nothing to indicate which signal corresponds to which meteorological element, and as the records may intersect, it might be thought that analysis would be difficult. In actual fact this is not so, and confusion very rarely arises. At the beginning of an ascent, knowledge of the ground conditions enables the three frequencies to be identified. In general the pressure frequency is the highest. At the top of the ascent, the humidity element always has the highest frequency. At intermediate points the character of the individual records is a guide. Pressure gives a smooth curve Temperature, while running roughly parallel to pressure, will have irregularities corresponding to inversions and changes in lapse rate. Humidity may make large changes.

Each recording chart lasts for ten minutes. At the conclusion of this period the completed record is handed over to a computer, who applies the necessary corrections, and from the calibration charts evaluates the frequencies in terms of pressure, temperature, and humidity. The results of the ascent are fully computed and ready for telegraphic transmission some 15 minutes after the balloon has burst.

There has been no attempt at automatic registration. While the labour of the operator would be eased by a mechanical recorder, the man himself could not be eliminated. As, also, such a recorder must necessarily be somewhat complicated, the reliability of the whole installation would be reduced and maintenance, always a difficulty in remote stations, increased.

The complete cycle of pressure, temperature and humidity measurements is repeated every 20 seconds during the earlier stages of an ascent, so with the average speed of ascent of 300 m./min., points every 100 metres in the atmosphere can be evaluated.

§ 3. CALIBRATION AND CONTROL

As the meteorological units are detachable from the main body of the instrument, these can be subjected alone to varying conditions in suitable chambers. The radio sonde proper stands outside the chamber and is connected to the units within by leads. It has been established that these leads introduce a negligibly small change in frequency, provided that the audio-frequency valve is fully decoupled with respect to radio frequency.

Pressure and temperature calibrations take place in the same vessel, which holds six units, pressure or temperature, at a time. The chamber is cooled by a bath of trichlo ethylene surrounding it. The bath is vigorously stirred and can be controlled in temperature by addition of solid CO_2 or by electric heating elements. The chamber itself is ventilated by a fan, and temperatures are measured at two points at the top and bottom by means of thermocouples. The use of two couples assures the operator that there are no undesirable temperature gradients at the moment of observation. In spite of forced air circulation, there is always some difference in temperature between top and bottom of the vessel, but this should not exceed 1° c. even at the lowest temperatures. This gradient is due principally to the fact that for constructional reasons the top of the chamber is not immersed in the bath.

The usual range of temperature calibration is from +25 to −70° C. Atmospheric temperatures as low as −85° C. are occasionally encountered in this country and −90° has been recorded in equatorial regions. A number of experimental calibrations, using liquid air as the cooling agent, have established that extrapolation of the normal calibration down to these temperatures will not produce errors greater than 1° C.

The pressure unit is given a full calibration from ground pressure to 50 mb. at 15° C. It is then cooled to −65° or −70° C. and readings for 300, 200 and 100 mb. taken, giving the frequency changes dF due to the change in temperature at these various pressures. As the variation of dF with pressure and temperature is complex, and due to several causes, it is not to be expected that all units will follow the law embodied in figure 7. It is found, however, that to a first approximation the behaviour of any unit can be represented by the curves shown, when the ordinates are multiplied by a constant factor Q_f, peculiar to that unit. It is therefore only necessary ideally to determine dF at one pressure and temperature in order to find Q_f. In practice the factor is determined at three pressure points, not only to improve the accuracy but to detect the small percentage of anomalous units which do not conform to the curves of figure 7. These latter are rejected.

This procedure for applying a temperature correction to the pressure unit gives only approximate results. Higher accuracy would undoubtedly be obtained by individual exploration of each unit. Practical considerations, however, rule this out. When large numbers of instruments have to be calibrated, the extra time involved would be prohibitive. Furthermore, the computation of a sounding must be as simple and rapid as possible if the results are to reach the forecasting centre in time to be of use, The approximate corrections lend themselves to swift computation by specially designed slide rules, embodying the data shown in figure 7. This would not be possible if individual correction charts were to be used with each instrument.

Humidity calibration is carried out in a separate apparatus. This comprises a chamber holding 24 units, in which the air is rapidly circulated, by means of a blower, past ventilated dry- and wet-bulb thermometers for measuring the humidity and through one of four vessels, containing respectively warm water, saturated solutions of sodium nitrate and calcium chloride, and silica gel. Any one vessel can be selected by means of a multiple valve, so that relative humidities of 100, 70, 40 and 10% are readily obtained.

Control corrections

Before the radio sonde is used in the air, the calibration at surface values of the meteorological variables is checked in a ventilated screen. These control readings are made with the instrument ready for flight, and take place a few minutes before release.

Slight changes in frequency relative to the calibration values are usually found. These arise from a variety of causes:—

(*a*) Differing standards of frequency at calibrating and observing stations.

(*b*) Use of different battery voltages during calibration and control.

(*c*) Influence of leads and of the surrounding metal chamber of the calibration apparatus.

(*d*) Secular changes in the meteorologically sensitive elements.

Errors due to (*a*) never exceed 0·25 c./s., as each station is equipped with an electrically maintained tuning fork for the standardization of frequency. Some variation in voltage occurs from battery to battery, which in spite of the stabilization introduced into the oscillating circuit may in extreme cases give rise to 0·5 c./s. change. The effect of (*c*) is relatively constant from instrument to instrument, and lies between 0·1 and 0·2 c./s. Thus although (*a*), (*b*) and (*c*) are all small, their cumulative effect may be of importance. Under (*d*) are included changes with time of the aneroid capsule and bimetallic strip, possible distortion of the frames due to mechanical shocks, and variation in the permeability of the mumetal owing to ageing or rough treatment in transport.

The overall change in frequency due to all causes amounts on the average to 1 or 2 c./s. A constant correction, called the control correction, is applied to the calibrations. This proceeding is somewhat arbitrary as it is not to be expected that the frequency changes remain the same at all values of the gap in the magnetic circuit. One factor, (*a*), is obviously independent of gap, and (*b*) and (*c*) give rise to errors which inversely vary as the gap, while (*d*) usually increases with gap. Thus it is impossible to lay down a general rule for variation of correction with frequency, as in any particular case the relative importance of the various factors is unknown.

The application of a constant correction is therefore only an approximation, but has the merit of allowing rapid computation. In order to limit the errors that may arise from this approximation, no radio sonde is used whose control corrections for pressure and temperature exceed 5 c./s.

§ 4. PERFORMANCE

It is not easy to assess the accuracy of a radio sonde. Direct comparison with a recording meteorograph attached to the same balloon will reveal gross errors, but the performance of the meteorograph itself is not sufficiently well known for critical work. Comparison against aircraft thermometers gives some information, but as it is not possible to ensure coincidence in place and time for both instruments, a lengthy series of measurements is required, and only mean figures for accuracy are thereby obtained. It is also not possible by this means to check the radio sonde at altitudes greater than 12 km., or about 200 mb. pressure. It is also generally impossible to separate pressure and temperature errors. But pressure measurements alone may be accurately assessed by comparison of heights computed from the readings with those directly determined by radar.

Accuracy of temperature measurements

Apart from the use of aircraft comparisons, casual errors may be assessed from the following experiments. A series of ascents was made with radio sondes in which the humidity units were replaced by extra thermometer units. Each unit of a pair was separately calibrated so that the final results include calibration errors, so far as these are not systematic.

The average difference in six flights between the pair of units in each instrument is given in table 3. In view of the fluctuations, the variation of the difference

Table 3

Presure level (mb.)	800	600	470	350	250	180	100	75	50
Difference (° c.)	0·45	0·50	0·67	1·00	0·56	0·61	0·22	0·44	1·00

with pressure is not significant. As the errors may be assumed to be equally distributed between the two elements, the probable error of one is ± 0°·4 c. In this figure are included all sources of casual error, except that due to variation of battery voltage during the ascent.

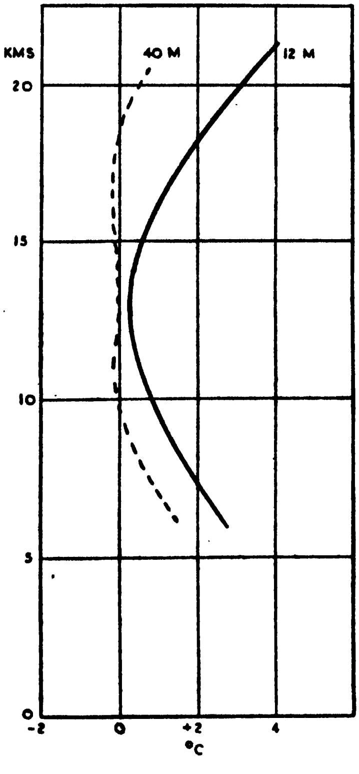


Figure 9. Difference of temperature between ascent and descent.
Full curve: Suspension between balloon and instrument 12 m.
Dotted curve: Suspension between balloon and instrument 40 m.

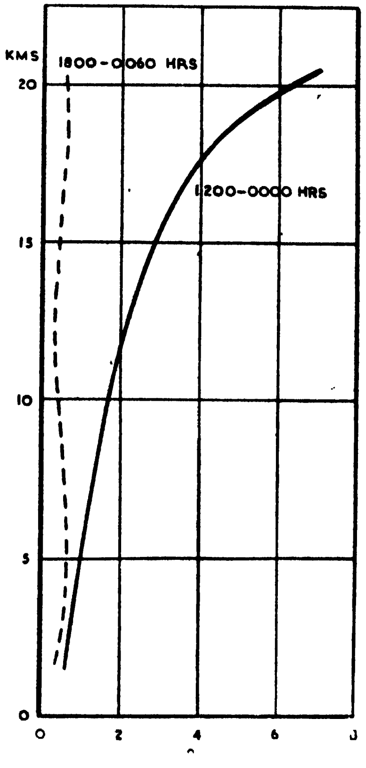


Figure 10. Mean difference of temperature between soundings at 1200 and 0000 hours (full curve), for May to July 1945, and between soundings at 1800 and 0600 hours (dotted curve).

The chief systematic errors arise from the effect of solar radiation. These are of two classes, respectively due to the action of the sun on the balloon and on the instrument itself. It is well known that solar radiation heats the balloon envelope to temperatures of which there is no reliable estimate but which may be 10 or 20° c. above that of the surrounding air. What has not been fully appreciated in the past is that the balloon as it ascends leaves a wake of heated air through which the radio sonde moves. The consequent temperature error becomes more important with increasing altitude, but can be minimized by using a sufficiently long suspension between balloon and instrument.

The error does not arise during descent, so that comparison of ascent and descent records will reveal its magnitude. Results of such a comparison are shown in figure 9. The full curve shows the mean difference between ascent and descent for a series of soundings around midday, when the distance between balloon and sonde is 12 m. The dotted curve gives the similar difference when a 40-m suspension is used. It is seen that with the shorter suspension the ascent may give temperatures too high by as much as 4°·0 c. at the highest levels, but that the error is negligible at a height of 14 km., or 150 mb. At lower levels it again becomes appreciable, but this is not due to radiation but to other factors, such as a tendency for the battery to run down towards the end of a flight, and to a pressure error, discussed below, which gives rise to a systematic difference in pressure recorded on ascent and descent. This pressure error is effective in falsifying temperatures only below the tropopause, as above this level temperatures do not vary with pressure. It will be observed that this apparent cooling on the descent at levels below the tropopause is also found with the long suspension. With a short suspension, the fall in temperature from the value on the ascent to that on the descent is nearly instantaneous and is a very pronounced feature of the flight record. When the supporting string is lengthened to 40 m. no sign of this sudden drop in temperature is found.

The effect of direct insolation of the instrument is more difficult to measure and to remove. It may be studied by comparing temperatures at the same height taken during neighbouring soundings at midday and midnight. The means taken over a long period should eliminate variations due to changing weather conditions. The full curve of figure 10 shows the mean difference in temperature at 0000 and 1200 hours during the three months May to July 1945, the observations being taken from three English stations. Such a curve shows not only the radiation error of the instrument but also any diurnal variation in the temperature of the air itself. It is indeed believed that the contribution of instrumental insolation is small, and that the diagram in fact gives a nearly true picture of the real daily temperature changes.

The evidence for this belief is as follows :—

(*a*) Many variations have been made in the form of the radiation shields around the temperature element. Also the surfaces of the bimetallic strips have been changed from a dull matt to a highly polished nickel-plated tape. This caused wide variation in the amount of radiation which the element was capable of absorbing. None of these changes has been found to make a significant change in the form of the curve in figure 10. It is reasonable to assume therefore that the instrument is insensitive to incident radiation.

(*b*) Observations are taken at 0600 and 1800 hours, and the temperature differences at these two times at various heights have been determined. A true diurnal variation will cause the air to be warmer at 1800 than at 0600 hours owing to the phase lag between air temperature and solar radiation. Direct radiation effects on the radio sonde will be equal at both times and will not appear in the difference. The results are shown by the dotted curve of figure 10. There is a difference, nearly constant with height, of about 0°·5 c. between the two times of observation, from which we can conclude that there is a true diurnal variation of temperature.

(*c*) The height of the radio sonde at any point can be computed from the pressure at that point and the temperature distribution below the instrument, using the usual barometric formula. The height can also be directly determined by radar methods to a much higher degree of accuracy. If the observed 1200–0000-hour differences (full curve of figure 10) are largely radiation errors, then in computing heights of a

> daylight sounding more accurate results would be obtained by utilizing the midnight rather than the midday temperatures. Piagsa (1946) has shown that this is not so. If the radar measurements are taken to be exact, the mean errors in height for about 50 daylight ascents are −86 metres when the temperatures as actually measured are used for computing, and +1000 metres when temperatures derived from the previous midnight soundings are taken.

We are therefore led to believe that instrumental errors due to radiation are small, and they can be roughly estimated as not exceeding 20% of the values given by the full curve of figure 10. At the very lowest pressures a larger contribution due to insolation cannot be excluded, as the number of observations at less than 60 mb. is much less than that at higher pressures.

A daily variation in the temperature of the upper air is not unexpected, though it is difficult to explain its magnitude. This question has been recently discussed by Dobson (1946).

Accuracy of pressure measurements

This can best be estimated by comparing the heights as computed from the radio-sonde observations with those directly determined by radar. It is routine practice at some stations to attach to the balloon a radar reflecting target and to observe its motion by means of a standard Army centimetric radar set, type A.A. No. 3 Mk. II. The high precision of range and elevation measurements gives an acceptable standard for checking the radio sonde.

Two series of comparisons have been made, by Harrison (1944) and by Piagsa (1945). The results are summarized in table 4, which gives not the actual height errors but the equivalent errors in pressure.

Table 4

Height interval (km.)	Radio-sonde pressure errors (mb.)	
	1944 series	1945 series
0 to 5	5·4	6·4
5 to 10	2·8	8·2
10 to 15	0·4	8·0
15 to 22	1·1	6·8

The two series made at different stations at different times are significantly different. The 1945 series is distinguished by the fact that the pressure control corrections were throughout larger than normal, for reasons not yet fully explained. These results are not therefore fully representative. The systematic errors in the lower layers in the 1944 series are partly explained by the use at that time of an incorrect temperature correction diagram. A more accurate diagram would reduce the errors up to 7 km. by about 2 mb.

Casual errors may be assessed in the same way as for temperature measurements. A series of ascents was made with instruments carrying two pressure units. The probable error of a single determination was found to rise from ± 1·5 mb. at 1 km. height to a steady value of ± 4·5 mb. from 13 km. upwards. The source of casual error is due to the application of the temperature correction. As previously explained, the curves in figure 7 are the means for a large number of instruments. Individual radio sondes will not follow the curves exactly. Systematic errors can arise not only when large control corrections are found, but also if the pressure unit is not at the temperature indicated by the thermometer. There is no systematic difference in pressure errors between night and day soundings, so that the radiation shielding of the pressure unit must be efficient, but as different parts of the unit make different contributions to the temperature effect and have different thermal lags, some error must arise from this cause.

This differential thermal lag is particularly apparent when heights on the ascent and descent are compared. The tropopause is found to be on the average at 10 mb. lower pressure on the descent. This is in the opposite direction from what would be expected from elastic hysteresis of the capsule, and can be attributed to the thermal lag of the mumetal core and coils of the inductance. This lag is of course only operative during the ascent, as on the descent the instrument traverses a region of practically constant temperature before the tropopause is reached. We should expect therefore that the descent should give a more accurate value of the pressure. The results shown in table 4 (1944 series), however, are not in accordance with this view, but indicate that the systematic error on the ascent is small. It is plain that there must exist another source of error, of approximately equal magnitude but opposite in sign, to that due to lag. There is not sufficient evidence to define its origin, but it seems likely that the correction curves in figure 7 are not sufficiently accurate. As they are obtained from only a small fraction of the total number of instruments used, some sampling error is to be expected.

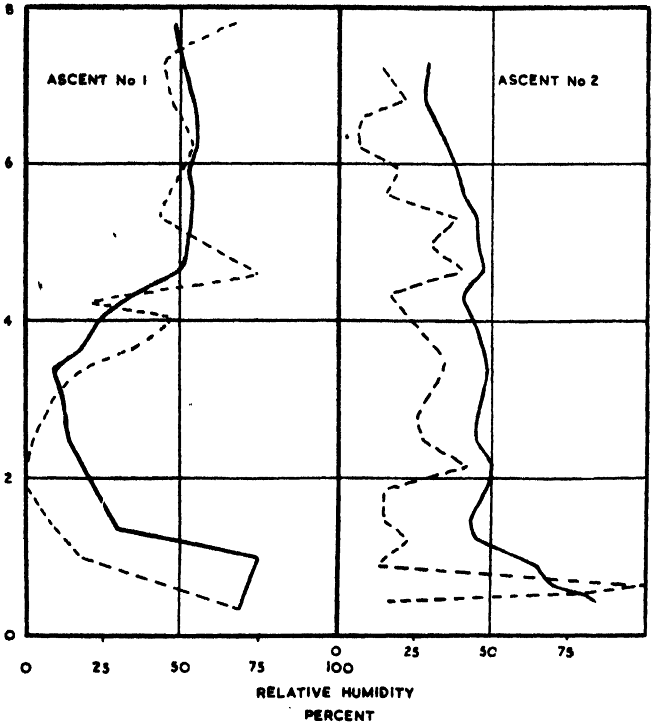


Figure 11. Comparison between humidity measurements by radio sonde (full curve) and Dobson and Brewer frost-point hygrometer on aircraft (dotted curve).

We conclude, therefore, that the pressure element is subject to casual errors of up to ± 5 mb., and while the systematic error does not in general exceed 2 mb., differences in manufacture of various batches may lead to considerably larger values if the several sources of error do not cancel.

Accuracy of humidity measurements

Here the only means of assessing systematic errors is by comparing with measurements made on aircraft ascents. Because of the low temperatures, the Dobson-Brewer frost-point hygrometer (Dobson, 1946) is the sole instrument that can be used as a standard, and it can only be used for upper-air work in an aircraft, as it requires a skilled operator.

The relative humidity of the atmosphere varies rapidly not only vertically but also horizontally. It is necessary therefore for the aircraft to keep close to the balloon during the ascent. This is very difficult to achieve in practice. As no aircraft with a sufficiently high rate of climb was available, in the trials it was necessary to operate the radio sonde at about a third of its normal ascensional velocity. The ventilation of the goldbeater's skin was accordingly deficient and the lag intensified. The result of two trials, by Harrison and Brewer (1944), are shown in figure 11. The sluggishness in response of the goldbeater's skin, especially at the higher levels, where the temperature is low, is evident. Deviations in the lowest layers between aircraft and radio sonde are to be ascribed to actual differences in the air, due to patches of clouds. The temperature differences indicated by aircraft and radio sonde on these ascents were within the probable errors and showed no anomalies in the region of gross differences in humidity.

Errors up to 25% relative humidity are shown by these trials. They do not provide a really fair indication of the performance of the radio sonde, for the reasons mentioned above, and it is to be expected that in proper conditions the instrument may attain an accuracy of better than 15% R.H.

Mutual comparison of two humidity elements on the same radio sonde, on the same lines as for pressure and temperature units, shows that the average difference is 5% R.H., with a maximum of 10%. The self-consistency of this type of element is therefore reasonably good, with a probable error of a single determination of ±2½% R.H.

On the other hand it can be seen that humidity measurements by radio sonde are not satisfactory. While a fair measure of accuracy is to be expected at levels in which the temperature exceeds −20° C., at lower temperatures the instrument gives little indication of the true conditions. No simple method of measuring humidity applicable to radio sondes will give acceptable results in this region. The reason for this lies in the exceedingly small quantity of water vapour which the air can contain at these low temperatures.

ACKNOWLEDGMENTS

The development of the Kew radio sonde owes much to many late colleagues in the Meteorological Office. Besides those referred to in the text, especial mention must be made of Dr. H. Carmichael, who was largely responsible for the design of the meteorological units, and of Messrs. L. G. H. Dines, I. M. Hunter, and A. J. Lander, who rendered much assistance in various stages of the work. Thanks are also due to officers in charge and to staff of the several radio sonde stations in the British Isles, who by their criticism and suggestions have done

much to improve the instrument, and whose keenness and attention is an important factor in establishing a high degree of precision.

I am indebted to the Director of the Meteorological Office for permission to publish this paper.

REFERENCES

Dobson, 1946. *Proc. Roy. Soc.*, A, **186**, 146.
Dunmore, 1939. *J. Res. Nat. Bur. Stand.*, **23**, 701.
Glückauf, 1947. *Proc. Phys. Soc.*, **59**, 344.
Harrison, 1944. *Meteorological Research Committee Reports*, MRP 203.
Harrison and Brewer, 1944. *Meteorological Research Committee Reports*, MRP 205.
Johnson, 1946. *Nature, Lond.*, **157**, 247.
Marth, 1944. Unpublished Report from the Marine Observatorium, Greifswald.
Piagsa, 1945. Unpublished Report from the Meteorological Office.
Piagsa, 1946. Unpublished Report from the Meteorological Office.
Thomas, 1938. *Proc. Roy. Soc.*, A, **167**, 227.
Väisälä, 1937. *Acta Soc. Sci. Fenn.*, **9**, 9.

THE ACCELERATION OF CHARGED PARTICLES TO VERY HIGH ENERGIES

By M. L. OLIPHANT, F.R.S., J. S. GOODEN and G. S. HIDE

MS. received 21 *March* 1947

ABSTRACT. More experimental information about the nature of the binding forces between nuclear constituents is necessary before an advance in fundamental nuclear physics can be achieved. By considering the type of information which would be most useful, the conclusion is reached that it necessary to have available protons of energies of about 1000 Mev. in order to carry out the necessary experiments. It is with a method of obtaining protons of this energy that this paper is concerned. An examination of the possibilities of achieving such high energy protons by the existing methods leads to a pessimistic conclusion, and a new method is suggested.

This new method, the synchrotron, is described in principle, and its advantages are outlined, a very important factor being its comparatively low cost. An accelerator of this type is being built at Birmingham University with a grant from the Department of Scientific and Industrial Research, and its design is considered in some detail. The magnet and its excitation form the greatest part of the apparatus in size and cost. Several alternative methods are suggested and discussed for both the magnet design and its method of excitation. An air-cored magnet is considered but rejected because of the very large mechanical forces involved and the precision required in positioning the conductors. As a result an iron-cored magnet has been chosen for construction. The excitation of the magnet is to be achieved by a d.c. motor-generator supplied with a fly-wheel. The requirements of the accelerating system, in which is included a radio frequency which changes by a ratio of about 1 ; 36 during the acceleration, are quite exacting. The methods by which it is hoped that these requirements will be met are outlined. The problems associated with injection and extraction of the particles receive some attention, and a schematic description of the proposed vacuum chamber is included.

When protons of energies greater than 10^{10} ev. are to be obtained by a synchrotron, the cost of the device becomes overwhelming and some alternative method will have to be suggested. The application of the synchrotron being built at Birmingham to accelerating electrons, is limited to achieving electron energies of about 300– 00 Mev. because of radiation losses.

§1. INTRODUCTION

FURTHER advance in fundamental nuclear physics is dependent upon an increase in experimental information about the nature of the binding forces between the nuclear constituents. The most obvious way to obtain this knowledge is to extend Rutherford's method of exploration of nuclei by bombardment with fast particles to much higher bombarding energies than have hitherto been available and to examine the laws of scattering of protons and neutrons in very energetic collisions with similar particles.

At the present time there is no real understanding of the forces between the elementary particles and, indeed, no satisfactory explanation of the existence of only a certain limited number of such particles, with very different masses, some electrically charged, others uncharged. (Peierls, 1946). The primary problem from the point of view of the physics of nuclei is that of the proton-neutron interaction. The mass of the neutron is greater than that of the proton for reasons which are not at all understood; energetically it should be possible for a neutron to transform spontaneously into a proton and electron, but this transformation has not yet been observed. Attempts to explain proton-neutron forces in terms of virtual creation in the immediate neighbourhood of the particles of pairs of electrons or mesons or of quanta are little more than assumptions that fields of particular forms exist round them. Such attempts have failed to explain the observed binding energies of nuclei. It is unlikely that substantial progress will be made by further guessing in this field of physics unless such guesses are guided by fresh experimental facts.

Primarily, interest must centre round the interactions at close distances of approach between the elementary particles, viz. protons, neutrons, electrons, mesons. It is probable that very energetic neutrons can be produced only by bombardment of matter with high energy protons. From the practical point of view it is therefore necessary that protons and electrons should be accelerated to energies which are as high as possible and their interactions with matter observed. The relative value of protons and electrons for this purpose is difficult to determine in advance. The great success of the cascade theory of shower production in cosmic radiation suggests that the interactions of nuclei with electrons are better understood than the interactions with heavy particles.

It would appear, then, that it is essential to produce protons with energies as high as possible and somewhat less necessary to accelerate electrons to comparable energies. It is important to be quite clear about the importance of accelerating protons, for the acceleration of electrons to energies of the order of $0{\cdot}5 \times 10^9$ ev. is so much simpler and cheaper that there is a great temptation to find excuses for accelerating electrons and for postponing the difficult problem of proton acceleration.

It is necessary that observations should be carried out at energies at least equivalent to the proper energy of a pair of the hypothetical nuclear mesons, and if it is assumed that these mesons have the same mass as the free mesons observed in cosmic radiation, particles are needed with energies above about 300 Me v. Since recoiling nucleons can carry away at least half of the initial energy it is probable that bombarding particles with energies above 600 Me v. are desirable. The total binding energy of nuclei of medium atomic weight is of the order of 1000 Me v. and the character of nuclear reactions is likely to change in this region of energies.

It is clear that a good target figure at which to aim in the development of new methods of acceleration is 1000 Me v. or more. Experience of other methods of acceleration suggests that if the maximum energy for which the equipment is devised is 1000 Me v., the maximum useable energy is likely to be lower. Thus an equipment designed for 1000 Me v. will be reasonably certain to deliver energies well above 600 Me v. without straining the apparatus. To settle some questions, particles with energies much greater than 1000 Me v. may be required, but it is probable that these higher energies will be obtained only after some experience in the region of 1000 Me v. In what follows a particular method for obtaining protons with energies above 10^9 ev. is described after some consideration of reasons for preferring this method to others which have been suggested.

§ 2. LIMITATIONS OF EXISTING METHODS

Acceleration methods may be divided broadly into two classes. In the first are all systems in which the particles are accelerated along straight paths; the second includes all methods in which a magnetic field is used to bend the particles during acceleration into spiral or circular orbits.

High-voltage methods belong to the first class. Such systems possess inherent stability of particle paths, provided the ordinary rules of electron-optics are observed, but they are limited to energies less than about 10 Me v. and are troublesome above 5 Me v. To this class also belong the so-called linear accelerator methods in which the particles are pushed along by a travelling wave moving at suitable velocity along a wave-guide, or in which they pass through a series of resonators where the phases of the fields are suitably adjusted. It is an inherent defect of linear accelerators that it does not appear possible to achieve directional focusing of the particles at the same time as phase stability. It is necessary to use external focusing methods, such as an axial magnetic field, which is difficult for large apparatus and high energies, or to use thin foils across the exit openings from the accelerating gaps in the manner proposed by Alvarez.* He has commenced the construction of a linear accelerator for protons in which the exit from the gaps is covered with thin beryllium foil. He hopes to gain an energy of 1 Me v. in each foot of the accelerating system so that an apparatus to produce protons of 1000 Me v. would be about 1000 feet in length. An enterprise of this sort is practicable only when a long building can be provided or where the equipment can be used out of doors. If it is successful this equipment may provide the simplest and cheapest form of accelerator capable of extension to almost unlimited energies. There are no difficulties due to radiation from the particles as they move in straight lines and no troublesome problems of injection or extraction of the particles. However, the engineering problems are formidable and the proposed solution to the focusing difficulty is not yet proven.

One advantage of these linear methods of acceleration is that they are applicable equally to all charged particles.

Apparatus in the second class includes the cyclotron, synchro-cyclotron, betatron and synchrotron. Heavy particles, such as α-particles, have been

---

* Private communication.

accelerated in the cyclotron to energies of 30–40 Me v. by Lawrence and his co-workers. The relativistic increase in mass of protons at energies above about 20 Me v. makes it extremely difficult and wasteful of electric power to go to higher energies by the straightforward cyclotron method. This difficulty has been overcome by introducing a change of frequency during the acceleration (McMillan, 1945). The synchro-cyclotron is an extremely successful apparatus and promises to become the standard equipment for acceleration of protons and other heavy particles to energies of a few hundred million electron volts. However, to produce protons with an energy of 10^9 ev., a magnet is required with a field of 15 000 gauss over a circular pole of radius 15 feet. Such a magnet would weigh more than 10 000 tons and would be extremely expensive to build and operate. A magnet of this order of size is under construction in U.S.A., to be financed from Government funds, but it is unlikely that similar equipment can ever be available in academic laboratories.

The induction accelerator, or betatron, of Wideroe (1928) and Kerst (1941), has been developed successfully for the acceleration of electrons, but considerations of cost and complexity render it unsuitable for the highest energies, while it cannot be used to accelerate heavy particles.

Various other attempts have been made to develop accelerating systems which can reach high energies, some of them employing resonance in a magnetic field, as those of Schwinger * and Veksler (1945), which use a combination of guiding field and linear accelerator or are modifications of the betratron, as Wasserab's "Wirbelrohr". However, none of these systems is very attractive, and they have not yet been either built or operated.

§ 3. THE SYNCHROTRON

In September 1943 one of us submitted to the Directorate of Atomic Energy in the Department of Scientific and Industrial Research, a proposal for the acceleration of electrons and protons by a new method to energies above 10^9 Me v. Subsequently, and independently, similar proposals were made by McMillan (1945) in U.S.A. and by Veksler (1945) in U.S.S.R. The name *synchrotron* was suggested by MacMillan. The essence of the new method is the conception of stable circulating orbits which increase in energy through a cyclotron type of resonant acceleration as a result of an adiabatic variation of the magnetic field, of the frequency of the accelerating electric field, or of both. The success of the synchro-cyclotron † afforded convincing proof of the validity of the general conceptions of the stability of the orbits for a system for the acceleration of heavy particles in which the frequency changes while the magnetic field remains constant. Goward and Barnes (1946) were able to demonstrate that electrons can be accelerated in a system where the radius of the orbit and the applied frequency of the electric field are constant but the magnetic field increases with time. There is a third system in which both frequency and magnetic field are varied during the acceleration. This system has been considered in detail by us and is now under construction. In what follows we give a general analysis of the proposed method and the considerations which have led to the designs adopted.

---

* Unpublished note.

† Private communications from Berkeley.

The principal practical aspect of the synchrotron method of acceleration is that for energies of the order of 10^9 ev. its cost is not prohibitive. This is due to the fact that since the orbital radius is constant, a narrow annular guiding magnetic field can be employed, so that the first cost of the magnet is much less than for a cylindrical field as used in the synchro-cyclotron. Much attention has been paid to the method of producing this field in an economical and satisfactory manner.

The essential data for the design of a synchrotron are the radius of the mean orbit, ρ, the maximum value of the magnetic field, H, and the rate of revolution, ν, of the particles in the orbit which determines the frequency of the accelerating voltage. These quantities are connected by the formulae:

$$W=\sqrt{H^2\rho^2c^2+E_0^2}-E_0, \qquad \ldots\ldots(1\,a)$$

$$\nu=\frac{c}{2\pi\rho}\sqrt{1-\left(\frac{E_0}{W+E_0}\right)^2}, \qquad \ldots\ldots(1\,b)$$

where W is the kinetic energy of the particles and E_0 is the self-energy, m_0c^2/e of the particles.

It is clear that in order to obtain high energies the maximum value of the product $H\rho$ must be large. We are concerned here with the design of a synchrotron to produce protons of energy greater than 10^9 ev., and in what follows we shall assume that W is to be $1{\cdot}3\times10^9$ ev.

The magnetic field must be so shaped that the orbits are stable (Kerst and Serber, 1941) and, in order to obtain an appreciable output in spite of inevitable radial and axial oscillations of the particles about the mean orbit, the width and depth of the annulus in which the particles move must not be too small. The magnetic field varies from almost zero to its maximum value during each cycle of acceleration, so that the construction must be such as to allow of A.C. operation. Thus the field can be generated in three ways; by using a system of conductors properly spaced and carrying appropriate currents: by using a ring-shaped laminated iron-cored magnet, with pole-pieces of the proper contour; or by shifting the magnetic flux to and from the annular space by purely electrical methods or by rotating or oscillating an electromagnet near magnetic circuits of proper design. In any case the cost of the magnet and its exciting circuits is the major item of expense and the choice of the magnet system determines all other parts of a synchrotron equipment.

§ 4. MAGNETIC FIELD AND RADIUS OF ORBIT

The variation of the energy ϵ stored in the magnetic field of a synchrotron, with the radius of the orbit, for a given ratio of gap dimensions (volume v) to radius, is given by

$$\epsilon=(H^2/8\pi)\,.\,v\sim\rho$$

for a fixed final energy of the beam produced. This magnetic field energy must be supplied by a source of electrical power, and the provision of this power represents the largest single item of expenditure. The above relation indicates that the cost of the power unit should decrease with radius and that it would pay to use the highest possible magnetic field. However, limits to H are set either by the saturation of iron or by the dimensions of, and forces upon, conductors where iron is not used.

(a) *Air-cored magnet.*—We have considered a system of conductors in the form shown in figure 1, where a current flows in one direction for conductors shown in open section, and in the reverse direction for conductors shown in solid section. If the system is straight and long compared with its diameter, a sinusoidal distribution of conductors gives a uniform field across the equator. If such a system is bent into a circular toroid the field increases across the toroid along a radius of the circle. To make the field fall off along the orbital radius, the sinusoidal distribution must be distorted, while, to enable the beam to be injected and ejected, the central conductor must be removed and suitable compensating conductors added as shown. The correct position of the conductors cannot be calculated, but a distribution can be chosen arbitrarily and the field calculated numerically. By a series of successive approximations a distribution of conductors was found which gave a reasonable approximation to the field required. However, it is found that the conductors must be placed and held very accurately in position.

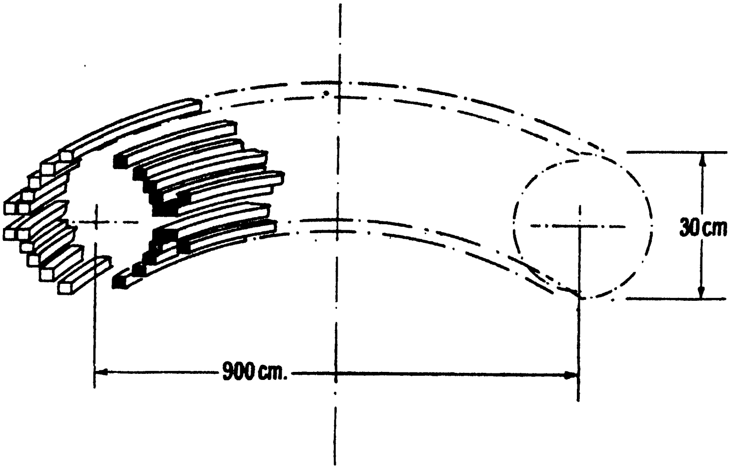


Figure 1. Arrangement of conductors for air-cored magnet.

The current in the conductors is independent of ρ for geometrical scaling, so that the relative cross-section of conductor increases as ρ decreases, giving departures from the field-form required. Also, as H increases, the forces on the conductors increase and the tolerances in position decrease, so that the problem of holding them in position rapidly becomes insuperable. Accordingly a compromise of 15 000 gauss, for which ρ is 450 cm., was chosen and the design of a synchrotron considered in detail. With a total of 22 conductors and diameter of orbital space of 30 cm., the allowable deviation of conductors from correct position is 0·1 cm., and a peak current of about 80 000 amperes is required. For operation at the equivalent of 25 cycles/sec. the peak driving potential across the coil is 25 600 volts. The forces on the conductors in this system are already of the order of the ultimate strength of copper conductors.

Continuous operation at 25 cycles would require a circulating energy of 2×10^6 kv.a. (6.4×10^6 w-sec.), while the copper losses would be 5×10^4 kw.

It is clear that such a system must be operated discontinuously by storing up energy continuously at a reasonable rate and discharging it at intervals through the coil. If the storage system is an electric condenser, a very bulky capacity battery of about 20000 microfarads is required, costing about £175000. The alternative of a short-circuit type of alternator has been considered in which the energy is stored as rotational kinetic energy, but the cost of the complete installation is of the same order of magnitude, while the engineering problems of installation and maintenance and the noise of such rapidly rotating machinery in an academic research laboratory render it even less attractive than the capacity battery.

(b) *Iron-cored magnet.*—The use of iron in a magnetic circuit reduces the volume of the useless magnetic field outside the orbital space and effects a saving of about a factor 2 in the energy stored. However, with a laminated structure the maximum flux density is limited to about 15000 gauss. The minimum dimensions of the orbital space to secure stability and a reasonable yield of particles are considered in another paper (Gooden, Jensen and Symonds, 1947). An iron-cored magnet has been designed in accordance with the results of these theoretical investigations and has the dimensions given in figure 2. The shape of the pole-tips has been found by model experiments in an electrolytic tank.

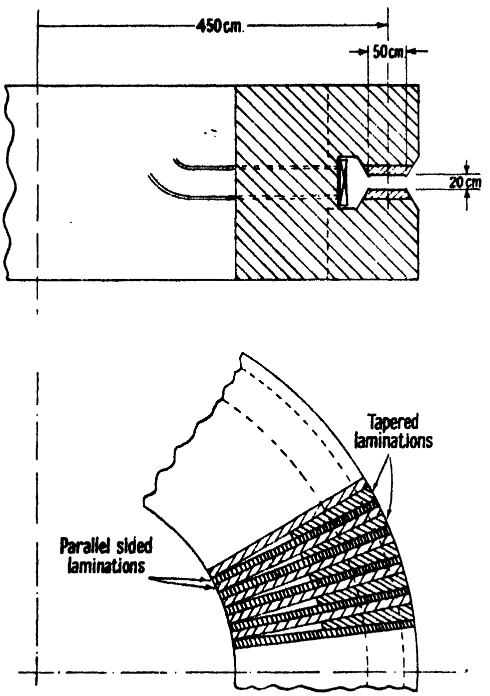


Figure 2. Proposed iron-cored magnet.

Operation of this magnet at the equivalent of 25 cycles would necessitate constructing it from laminated electrical steel which is in very short supply. Theoretical investigations (Gooden, Jensen and Symonds, 1947) indicated that the yield of protons might be improved by using a more slowly rising magnetic field and that the increased intensity of output in each pulse might compensate for a lower repetition rate. By improving the vacuum conditions it is thought that loss of particles by scattering in the residual gas can be reduced to an extent where a time of acceleration of about 1 second is practicable, with an initial injection energy of 300000 ev. A slow rate of change of field also has the advantage that the corresponding change in frequency of the accelerating potential can be more easily achieved by mechanical methods. Although the energy which must be supplied to create the magnetic field is the same as for shorter periods of excitation the power is reduced in proportion to the time of rise of field. In particular, for a time of rise of field of about 1 second, it becomes practicable to supply the energy from a d.c. generator which is provided with a flywheel in the manner used for "field-forcing" in switch-gear testing equipment.

At these very low frequencies the thickness of lamination which can be employed is large. The field is sensibly in phase over the orbital space for laminations 0·5 inch in thickness. It is thus practicable to build the magnet from rolled sheets of low-carbon steel and a very good space factor can be secured if some of these sheets are tapered. Through the cooperation of Sir A. McCance, F.R.S., of Colville's these special sheets are now in process of manufacture.

(c) *The electrical circuit.*—The relevant electrical data for the magnet are given in the following table:

Turns in winding	22
Cross-section of conductor	7 sq. cm.
Resistance	0·014 ohm.
Inductance	0·1 H.
Time-constant	7 sec.
Peak current for 15,000 gauss	11 000 amp.
Volts to give rise in 0·8 secs.	1100 volt.
Number of cycles of excitation per minute	6

The proposed cycle of operation is shown in figure 3. The generator, consisting of twin-coupled d.c. generators in parallel, driven by a 1500 h.p. motor and provided with a 36-ton flywheel, will be supplied by Messrs. Parsons, whose help we are glad to acknowledge.

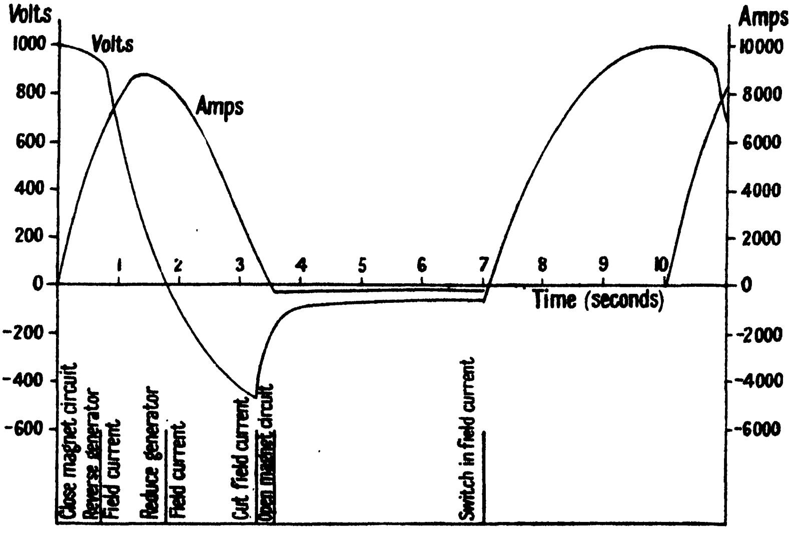


Figure 3. Cycle of operation of magnet and generator circuit.

§ 5. THE ACCELERATING SYSTEM

For the magnet under consideration the fundamental synchrotron equations (1 *a*, 1 *b*) become:

$$W = 300\sqrt{H^2(2{\cdot}03\times10^5)+(9{\cdot}61\times10^{12})}-9{\cdot}3\times10^8, \qquad \ldots\ldots(2a)$$

$$\nu = 1{\cdot}06\times10^7\sqrt{1-\left(\frac{9{\cdot}3\times10^8}{W+9{\cdot}3\times10^8}\right)^2}, \qquad \ldots\ldots(2b)$$

where *W* is in electron-volts and *H* is in gauss.

It is clear from figure 3 that H will increase approximately linearly. The values of ν for corresponding values of H determine the frequency of the accelerating voltage applied to the accelerating electrodes. For the slow rate of acceleration chosen, the energy added per revolution of the particles is only about 200 volts and the voltage amplitude of the a.c. applied to the electrodes need be of the order of only 1000 volts. There is an optimum value for this applied voltage which leads to greatest orbital stability and maximum output current, the reason for which is discussed in another paper. The same theoretical reasoning shows that there are no advantages to be gained by using more than one electrode, the frequency applied to which is $\nu_e = \nu$. It can be shown (Gooden, Jensen and Symonds, 1947) that ν_e must equal ν to within about 0·1% over the first one-hundredth part of the acceleration, and must not differ from ν by more than about 1% thereafter.

If protons are injected at 0·3 Me v., which is about the maximum for an internal "gun", then ν_e must vary from about 0·27 Mc. to 9·5 Mc., i.e. by a factor of more than 30. A circuit of low Q, which is tuned to about 1·5 Mc., will give an adequate response when driven over the lower frequency range. A factor in frequency of about 8, which remains, can be obtained by mechanical tuning of the relatively low Q circuit through a cam of suitable shape, the frequency being adjusted to the exact value required by electronic "pulling" of the oscillator produced by the magnetic field itself. It is particularly important that the frequency shall be correct at the time of injection which corresponds with a magnetic field of about 170 gauss. A detailed account of the high frequency system of this synchrotron will be given elsewhere.

§ 6. THE INJECTION SYSTEM

If a reasonable output is to be obtained from a synchrotron it is essential that as many particles as possible should be injected during the "acceptance" period of the cycle. As with the betatron the mode of injection and the subsequent motions of the particles must be so designed as to ensure that the protons do not collide with the gun system during subsequent revolutions. Besides the particle oscillations which occur in the betatron, there are further radial oscillations in the case of the synchrotron, which are associated with the phase oscillations. This problem is subject to analysis (Gooden, Jensen and Symonds), though it may be that some factors have been neglected and the conclusions from the analysis may be no more applicable than similar calculations made for the betatron by Kerst and Serber (1941). However, as a result of the analysis it had been decided to place the ion source and initial accelerating system above the orbital plane and to apply a vertical electric field which will make the orbits spiral downward during injection at a rate sufficient to ensure that they miss the gun after the first revolution. Uncertainties in the analysis of the initial motions of the ions may mean that this system will need considerable modification before the maximum output is obtained. Observation of the paths of the protons after injection with steady magnetic fields, and of the paths of alpha-particles from radioactive sources with fixed magnetic fields, should make it possible to correct the injection conditions, whether arising from position of the gun or the inevitable deviations of the magnetic field from the correct form due to inhomogeneous properties of the iron at these low magnetizations.

§7. THE EXTRACTION SYSTEM

In order to ensure that the protons are obtained in a definite beam, they must be deflected by an electric or magnetic field during a single revolution. The problem is much simpler than with the betatron or with small synchrotrons, as the period of revolution in the orbit at the maximum energy is much greater, owing to the large radius of the path. A deflecting voltage can be applied to a relatively long electrode in a time short compared with the time of revolution in the orbit (10^{-7} sec.), by connecting it to a large capacity through a spark gap which is triggered to break down at the end of the acceleration cycle, the electrode circuit being suitably damped to prevent oscillations. For instance, a field of 10^5 volts/cm. applied by an electrode 300 cm. in length would produce a deflection of about 5 cm. in particles of enegy 10^9 ev. Such an electric deflection might be combined with a magnetic shielding channel such as that described by Skaggs and others (1946) for use with the betatron, especially as the long time of acceleration would permit the mechanical insertion of such a channel after the orbits had settled down, thus avoiding the possible damaging disturbance of the orbits due to distortion of the magnetic field during the injection period, when the magnetic field is small.

§8. THE VACUUM CHAMBER

Acceleration systems which employ a varying magnetic field necessitate use of a vacuum chamber, the walls of which cannot carry appreciable eddy currents which would upset the phase and shape of the field. The so-called "dough-nut" in the betatron is made of glass or ceramic and is coated on the inner surface with a thin shielding layer of silver or other metal, but this is not a practical solution for a large synchrotron. Figure 4 gives a schematic view of the chamber proposed for the apparatus in Birmingham. Models are under construction and the final chamber may differ in detail from this.

The acceleration space is formed of corrugated strips of stainless steel, the widths of which are small enough to reduce the effects of eddy currents to small proportions. The strips, which are radially disposed, are fixed rigidly to the outer octagonal section, also of stainless steel, and are not in electrical contact except at this junction. The whole is rendered air-tight by stretching a sheet of non-porous rubber over the strips and clamping this tightly to the octagon. The rubber is shielded completely from the beam by the interlocking corrugations, and since these are short compared with the wavelength of the radio-frequency accelerating field, the rubber is not subject to appreciable high frequency fields. The flat faces of the octagon are closed with plates which carry the pump manifolds, exit port, the insulated leads to the accelerating electrode, the source of protons, deflecting electrode etc. The chamber is made in eight sections bolted together with suitable rubber gaskets in each joint, and exhausted by six oil-diffusion pumps each 15 inches in diameter, in the manifolds of each of which is a liquid air trap.

The accelerating electrode is laminated to eliminate the effect of eddy currents. The advantage of this construction for the chamber is that it gives easy access to the interior and allows of modifications to the electrode system etc. without removing the vacuum system from the magnet. Such flexibility is important in equipment which is experimental in design, and where considerable modification may be required as a result of practical experience.

§ 9. SYNCHROTRONS FOR HIGHER ENERGIES

The synchrotron can be extended to higher energies by increasing the radius of the orbit and scaling up the other dimensions, keeping the same magnetic cycle. Figure 5 shows the way in which the cost, power demand and radius vary with the energy. It is seen that while equipment may be built for energies of 2–3 × 10^9 ev. it would be prohibitively expensive to construct a synchrotron for 10^{10} ev., at any rate in Great Britain. If higher energies are needed another method of acceleration must be used.

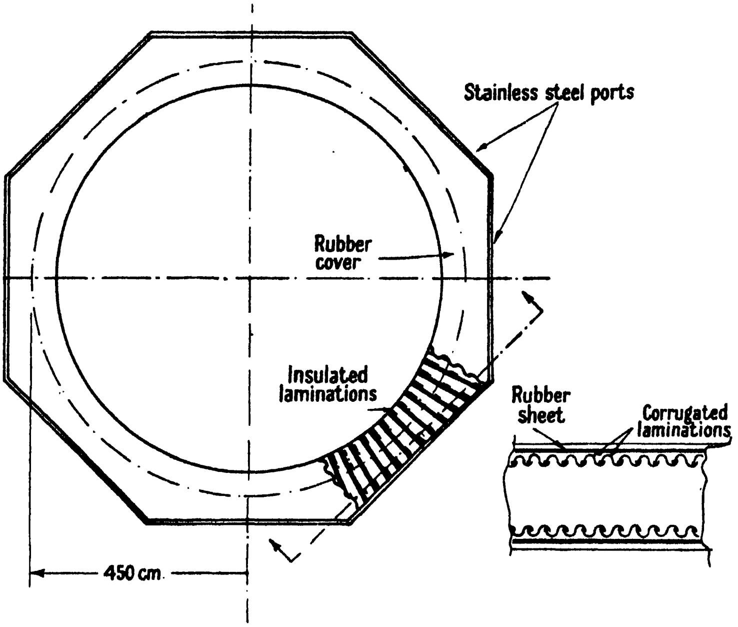


Figure 4. Schematic view of vacuum chamber.

§ 10. ACCELERATION OF ELECTRONS

It is impossible that energies as high as 10^9 ev. can be reached with electrons in a synchrotron of this type because of the excessive loss of energy as radiation by the particles due to their motion in a circle. This radiation loss can be calculated (Schwinger, 1945 and Schiff, 1946) and at 10^9 ev., with $\rho = 450$ cm., it is of the order of 20000 ev. per revolution. Thus to obtain electrons with energies comparable with those which can be reached with protons, the particles should be accelerated much more continuously, i.e. a large number of accelerating gaps would be needed with a voltage across each which is high compared with the nett rate of gain of energy, and driven at a correspondingly higher frequency.

It is difficult to estimate the maximum electron energy which could be obtained by operating the present equipment with a constan telectrode frequency of about 10 Mc., but it is probably in the region of 300–400 Me v.

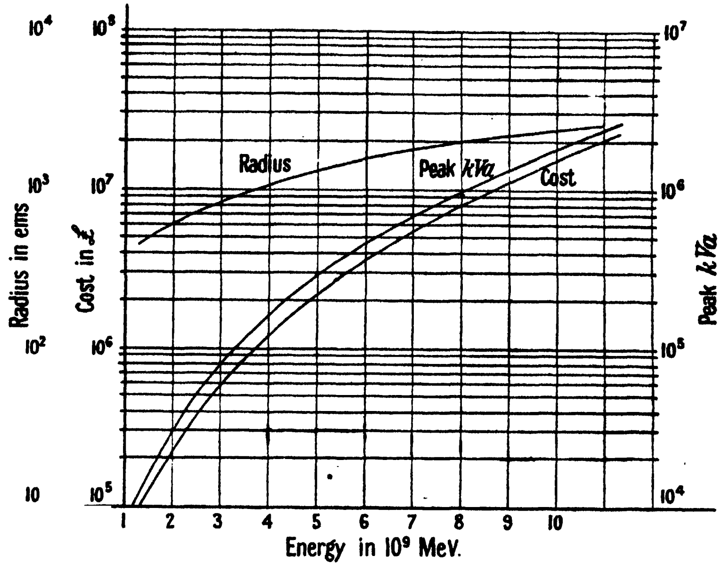


Figure 5. Variation of cost power demand and radius with particle energy.

REFERENCES

Gooden, Jensen and Symonds, 1947. *Proc. Phys. Soc.*, **59**, 677.
Goward and Barnes, 1946. *Nature, Lond.*, **158**, 413.
Kerst, 1941. *Phys. Rev.*, **60**, 47.
Kerst and Serber, 1941. *Phys. Rev.*, **60**, 53.
McMillan, 1945. *Phys. Rev.*, **68**, 143.
Peierls, 1946. *Nature, Lond.*, **158**, 773.
Schiff, 1946. *Rev. Sci. Instrum.*, **17**, 6.
Skaggs, Almy, Kerst and Lanze, 1946. *Phys. Rev.*, **70**, 95.
Veksler, 1945. *J. Phys. U.S.S.R.*, **9**, no. 3.
Wasserab. Unpublished.
Wideroe, 1928. *Arch. Electrotech.*, **21**, 387.

THEORY OF THE PROTON SYNCHROTRON

By J. S. GOODEN, H. H. JENSEN and J. L. SYMONDS

Read 20 *December* 1946 ; *MS. received* 21 *March* 1947

ABSTRACT. In the type of synchrotron for accelerating protons, the particle velocity, and consequently the accelerating radio-frequency, increase with increasing particle energy. In such a case the particle motion acquires properties which necessitate a careful control of some of the physical variables. In particular, it is found that within the stable limits of phase, non-relativistic particles, to a first approximation, possess undamped phase oscillations. The particles can be accelerated only so long as it is ensured that any factors affecting the phase oscillation amplitude are sufficiently small. It is necessary, therefore, to consider in some detail the physics of these oscillations, and in particular, of their damping.

It is found that there are no less than eight significant forces which can affect the behaviour of the phase oscillation amplitude. Four of these forces can be adjusted to some extent, the limitations being those of a practical nature. Thus it should be possible to accelerate protons in a synchrotron, if reasonable care is taken.

The problems of injecting the particles into a synchrotron working as such are considered. In this connection the radial oscillations accompanying the phase oscillations are described, and from this knowledge of the motion the time interval of injection is determined.

Numerical data and graphs illustrating the results are given for the case of the Birmingham synchrotron. Attention is focused throughout on the physical description of the motions, but detailed mathematical results are included.

§ 1. INTRODUCTION

General

The proposal for accelerating extreme relativistic particles (electrons) by the synchrotron was put forward by Veksler (1945), McMillan (1945) and Oliphant (1947) and has subsequently received considerable theoretical attention from many authors (Bohm and Foldy, 1946; Dennison and Berlin, 1946; Frank, 1946). In such a device, the particles (electrons) are moving with a velocity close enough to that of light to be considered as sensibly constant. In this case the properties of the particle motion make it unnecessary to control stringently most of the physical variables of the system. In particular, it is found that for a large range of phases, the phase oscillation amplitudes decrease moderately rapidly with increasing particle energy, and thus the factors affecting the phase oscillation amplitude do not have to be carefully controlled. It is convenient to use the principle of the betatron to accelerate the electrons to velocities sufficiently close to that of light, before the synchrotron operation is commenced (Pollock, 1946). Then initial injection problems are identical with those of the betatron (Kerst and Serber, 1941).

In the type of synchrotron suggested by Oliphant and McMillan for accelerating protons, the particle velocity, and consequently the accelerating radio frequency, increase with particle energy. In such a case the particle motion acquires properties which necessitate a much more careful control of some of the physical variables. In particular, it is found that within the stable limits of phase (see below) the particles, to a first approximation, undergo undamped phase oscillations. If the phase oscillation amplitude is allowed to increase, the phases eventually reach the unstable region and the particles are lost. The particles can be accelerated only so long as any factors affecting the phase oscillation amplitude are sufficiently small. It is thus imperative to have a more thorough understanding of the physics of the phase oscillations and in particular of factors causing variations in the phase oscillation amplitude.

Initial acceleration by a betraton action is not practicable with protons. It is not possible to inject protons with energies high enough for them to be accelerated only in the relativistic range of velocities where they are inherently stable. It therefore becomes necessary to investigate the motions of particles following their injection. In what follows these problems are considered for the Birmingham synchrotron (Oliphant, Gooden and Hide, 1947).

Basic description of phase oscillations

The phase oscillations occurring in a synchrotron accelerating extreme-relativistic particles have been described by Veksler (1945) and McMillan (1945). Since the physical properties of these oscillations constitute the most part of what follows, it will be convenient to mention here in detail some of the essential features of such an oscillation. The variation of the amplitude of the oscillations and other refinements are treated in later sections.

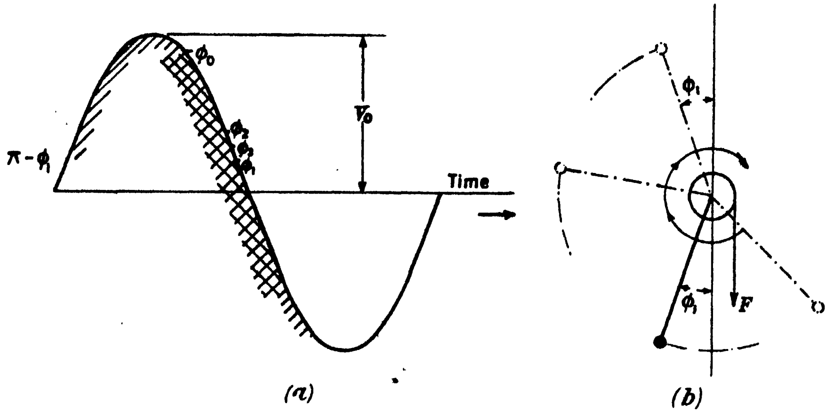


Figure 1. (*a*) Accelerating voltage amplitude to illustrate phase oscillations. (*b*) Pendulum analogue.

The basic action of the phase oscillations can be seen by reference to figures 1 and 2. It can be shown that the number of accelerating gaps is of no significance to the argument which follows and so, for simplicity, a one-gap system is assumed throughout. Consider firstly the case of extreme-relativistic particles, and assume the particles travel in circular orbits. The voltage amplitude, applied across the R.F. accelerating gap, is made greater than the energy (in volts) required to be added to a particle per revolution, in order to maintain it on the central orbit. The R.F. is so chosen that a particle moving on this orbit will always arrive at the gap at a certain constant R.F. phase. This phase is ϕ_1, so that the energy to be added per revolution to maintain the particle on the central orbit is $eV_0 \sin \phi_1$. This central orbit is called the stable orbit. A particle moving instantaneously

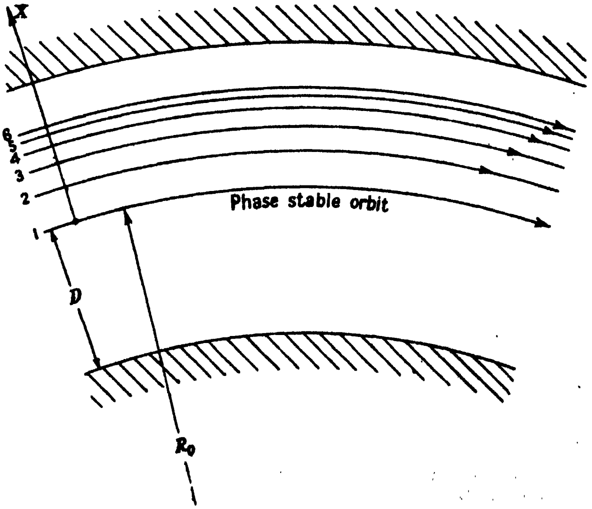


Figure 2. Illustrating the progressive increase in radius of a particle undergoing phase oscillations.

on the stable orbit but arriving at the gap at a phase ϕ_0, say, receives extra energy and will thus increase its radius. This means that now the particle will lose in phase each revolution because of its larger circular path. It will therefore change its phase on next arriving at the gap in the direction of ϕ_1, gain excess energy and further increase its radius. This process will continue until the particle reaches an orbital radius and a phase such that the energy gain per rev. is just sufficient to maintain it on its orbit. Then, because of its larger circular path, it still will be losing phase and, gaining less energy than necessary to maintain it in its orbit, will decrease its orbital radius so that the whole process is reversed.

This action, when continued, sets up the phase oscillation and its accompanying radial oscillation. Several useful relations should be stated here. (Notation is given in the appendix.)

(1) The change of phase per rev. (called "phase velocity") is directly proportional to the difference between the orbital radius and that of the stable orbit, i.e. $\frac{\partial\phi}{\partial q}\propto \delta R$, where q is the number of revolutions undergone by a particle starting from rest.

(2) The change in radius per revolution of the particle (rate of change of radius) is directly proportional to the ratio of the excess energy received by the particle per rev. to the total energy of the particle, i.e.

$$\frac{\partial R}{\partial q}\propto\frac{eV_0(\sin\phi-\sin\phi_1)}{E}.$$

(3) It follows that the acceleration of phase $\left(\frac{\partial^2\phi}{\partial q^2}\right)$ is proportional to $\frac{\partial R}{\partial q}$ and consequently $\frac{\partial^2\phi}{\partial q^2}$ varies as $\frac{eV_0(\sin\phi-\sin\phi_1)}{E}$. Thus, as was first pointed out by McMillan (1945), the phase motion for a given particle energy is identical in form with the angular motion of a simple pendulum under the additional influence of a constant torque so that its stable equilibrium position is ϕ_1 (see Figure 1 b). There is also an unstable position at $\pi-\phi_1$. This analogy is of considerable value and will be used quantitatively in another section.

Thus there is a range of phases within which the phase can oscillate stably. This range is bounded on the one side by the phase $(\pi-\phi_1)$ and on the other by the phase ϕ_2, which is given by the relation $(\pi-\phi_1+\phi_2)\sin\phi_1+\cos\phi_2=1-\cos\phi_1$. As in the pendulum case, the oscillations about ϕ_1 and the corresponding radial oscillations of the particle in the synchrotron will be asymmetrical. If the phase of the particle exceeds the stable limits it will cease to oscillate and will increase continually. The particle will then spiral inwards and hit the inside wall of the synchrotron. This motion corresponds to the pendulum swinging continuously in a circle under the action of the constant torque when the bob is placed outside the stable limits.

In what follows reference will be made to the " phase restoring force " which, on the above argument (see point (3) above), will be proportional to

$$\frac{eV_0(\sin\phi-\sin\phi_1)}{E}.$$

All arguments concerning the conservation of energy of a pendulum have their valid analogy in the phase oscillation case.

When the particle is non-relativistic, and the magnetic field is uniform, the cyclotron conditions hold and there can be no phase stability. Furthermore, because of the increased orbital radius of a particle which receives excess energy, the particle will enclose a larger amount of changing magnetic flux, be further accelerated on this account and thus eventually hit the synchrotron walls.

If now the magnetic field is made to decrease radially outwards (a condition which is necessary to ensure vertical stability) the particle, on gaining excess energy (and velocity), will increase its radius more than when in a uniform magnetic field. The extra path thus introduced will cause the particle to change its phase as it rotates, just as in the extreme-relativistic case. Thus phase stability is introduced. The effect of the induction forces is now to increase the amplitude of the radial oscillations accompanying the phase oscillations, but they cannot prevent the oscillations from occurring.

Thus there is little essential difference, in principle, between the non-relativistic and extreme-relativistic cases. However, large differences occur in the damping of the phase oscillations.

§2. INJECTION

Particle motions

A particle injected into the synchrotron has, in general, two resulting radial oscillations. One is identical with the radial component of the oscillation occurring in a betatron and is described by Kerst and Serber (1941). This will be termed the " injection oscillation". The other radial oscillation is that accompanying the phase oscillation and has just been described.

In order to obtain a picture of the injection process, consider particles of uniform energy being directed continuously into the synchrotron. Neglect at first the injection oscillations. Then every particle will experience the phase and radial oscillations as described in the previous section, provided it enters the gap during the stable region of R.F. phase. Of these, only particles with maximum radial oscillation amplitudes less than the half width of the accelerating chamber will continue their motion and be accelerated. At a certain time, during a certain R.F. cycle, the particles entering will have their instantaneous orbit coinciding with the central stable orbit. For convenience this R.F. cycle is called the *zero*th R.F. cycle. Particles arriving during earlier or later R.F. cycles will have instantaneous orbits greater or less, respectively, than the stable orbit. These R.F. cycles are called the $-u$th and $+u$th respectively.

Consider particles entering during the stable phase range of the uth R.F. cycle. They will form a long bunch slightly inclined to the instantaneous orbit of the particle entering at the phase ϕ_1 (see figure 3). Because all the particles are moving on an orbit smaller in radius than the stable orbit, they will all have the same initial phase velocity and their phases will move up the voltage curve of figure 1. The particle arriving at phase ϕ_1 will just begin to gain excess energy and will therefore begin to increase its orbital radius. Those particles in front of the bunch, receiving much more than the stable energy, will increase their orbital radii rapidly, while conversely those at the back of the bunch will decrease their orbital radii, but more slowly. Consider the point 0 which moves along the stable orbit with the phase velocity of the R.F. Then the bunch of particles, as a whole, will rotate around 0, roughly remaining tangent to the curve traced out by the particle entering at phase ϕ_1. Successive stages of this motion are shown in figure 3. It is readily seen what determines the length of bunch which can be accelerated, and consequently the accepting phase range for the *u*th R.F. cycle. This is modified by the injection oscillations. The curve traced out by any particle is in general not simple, but for small amplitudes and for a rectangular reference system it becomes an ellipse.

If now the injection oscillations are also considered, the resulting motion of a particle injected during the *u*th R.F. cycle will be as shown in figure 4. The length of bunch accepted during a given R.F. cycle is given by the intercept of the instantaneous orbit of the particle arriving at phase ϕ_1 on the mean curve traced out by the particle which just misses the walls of the chamber. This determines the phase range of acceptance. The number of R.F. cycles accepting particles is obtained from this picture by determining when the instantaneous orbit added to the injection oscillation amplitude just touches the synchrotron chamber walls (or unstable region of magnetic field). Because of the random phases of the various oscillations among the particles from all R.F. cycles, the bunches will mix together to give a large resultant bunch of roughly uniform density and of a shape given by the envelope of the particle motion shown in figure 4.

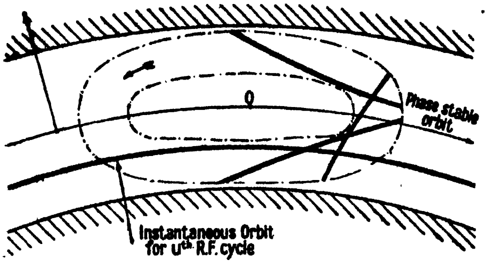


Figure 3. Successive stages of the motion of a bunch of particles entering during one R.F. cycle.

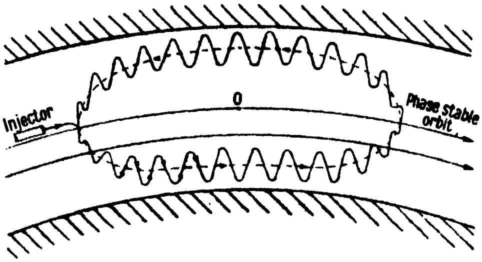


Figure 4. Motion of a particle injected into the synchrotron relative to a stable particle.

For a non-relativistic particle (as will be shown later) this big bunch will decrease in width as (kinetic energy)$^{-1/2}$ but will not decrease in length. For the Birmingham proton synchrotron the decrease in length, due to the relativistic phase damping introduced in the latter half of the acceleration, is only about four times, whereas the width of the bunch is reduced to a few millimetres. For an extreme-relativistic electron accelerator, the decrease in length is proportional to (total energy)$^{-1/4}$ and the width decreases more rapidly than (total energy)$^{-5/4}$.

Summing the number of R.F. cycles and the time intervals per R.F. cycle over which the particles are accepted, gives the total effective time interval, *T*, during which particles entering the synchrotron are eventually accelerated. For the type of injection system proposed for the Birmingham synchrotron

(Oliphant, Gooden and Hide, 1947), this time interval will be of paramount importance in determining the number of particles which will be accelerated.

Factors affecting the time interval of injection

It has been seen that the effective time interval of injection is made up of two parts: (*a*) the number of R.F. cycles accepting particles, (*b*) the phase ranges for each R.F. cycle over which particles can be accepted. These permissible phase ranges are limited by two conditions. The first limitation is that the R.F. phase at which a particle arrives at the gap must always be inside the stable phase range discussed in section 1. The second limitation is that the radial oscillation arising from the phase oscillation must always lie inside the synchrotron. These two limitations, in general, determine the permissible phase range for each R.F. cycle. It is evident that this phase range will be a maximum when the two ranges are made identical by adjusting the voltage amplitude on the accelerating electrodes (see below).

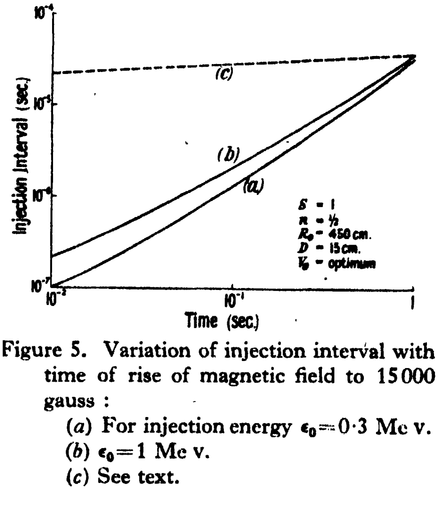


Figure 5. Variation of injection interval with time of rise of magnetic field to 15000 gauss :
(*a*) For injection energy $\epsilon_0 = 0{\cdot}3$ Me v.
(*b*) $\epsilon_0 = 1$ Me v.
(*c*) See text.

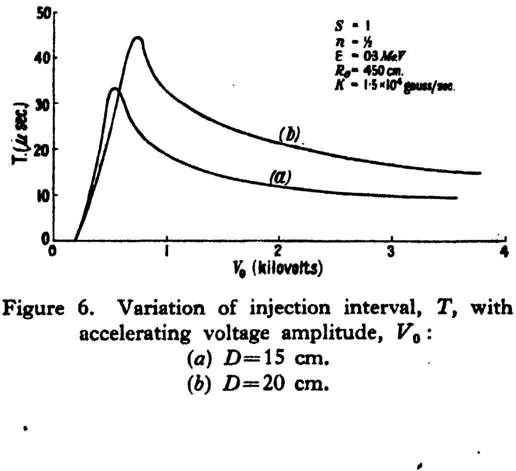


Figure 6. Variation of injection interval, T, with accelerating voltage amplitude, V_0:
(*a*) $D = 15$ cm.
(*b*) $D = 20$ cm.

Both the number of accepting R.F. cycles and the accepting time intervals per R.F. cycle are influenced by a large number of physical variables. The most important of these are the rate of change of magnetic field, the initial energy, the voltage amplitude applied to the accelerating electrode and the width and depth of the accelerating chamber. The influence of such variables in relation to design is now considered in detail.

Rate of rise of magnetic field

Figure 5 illustrates the advantage to be gained by decreasing the rate of rise of the magnetic field. The conditions are those of optimum voltage amplitude, so that the increase in T obtained represents an increase of the optimum values. The corresponding voltages are also shown. The reasons for this behaviour of the effective time interval T are two-fold. Firstly, as $\partial B/\partial t$ is reduced, the time taken for the field to change the amount necessary to bring the instantaneous orbits from one side of the accelerating chamber to the other is increased proportionately to $(\partial B/\partial t)^{-1}$, and the time interval accordingly is increased. Secondly, as $\partial B/\partial t$ is decreased, the energy to be added per revolution is decreased. This means that the voltage amplitude must also be reduced in order to regain the optimum phase range. Because of this reduced amplitude, the excess energy received by a particle at a given phase will be less than before. Thus particles arriving at a given phase (same initial energy) will not increase their radii as much as before. This means that particles can be accepted over a larger phase range per R.F. cycle. The voltage is then reduced sufficiently to optimize T, thus making the increased phase range available. Obviously there is a limit to the increase gained in this way as ϕ_1 approaches zero. The dotted curve in figure 5 shows the expected mean current determined on this basis for a continuously oscillating magnetic field as a function of the oscillating frequency.

Initial energy

The injection energy (ϵ_0) influences the effective time interval of injection T in three ways, the total effect being that T increases roughly as the square root of the initial energy. This behaviour is illustrated in figure 5 where the curves of T against $\partial B/\partial t$ are plotted for several values of the initial energy. It is thus best to design for as high an injection energy as is practical for the system proposed. There are other advantages to be gained by choosing a high initial energy such as a reduction in the radio frequency change necessary for a proton synchrotron.

The three effects of the injection energy, ϵ_0, on T are:

(1) The greater the initial energy, the greater the initial velocity, and hence the higher the frequency of the initial R.F. This means the time intervals of all given phase ranges are reduced in the proportion $(\epsilon_0)^{-\frac{1}{2}}$.

(2) As ϵ_0 increases, so does the magnetic field required to give the same orbital radius. A greater absolute change in the magnetic field is needed to move the instantaneous orbit of the particle from one side of the synchrotron chamber to the other. The rate of change of magnetic field is the same, so that a larger time will elapse for this process $(\propto \epsilon_0^{\frac{1}{2}})$ and consequently the number of accepting cycles is increased $(\propto \epsilon_0^{\frac{1}{2}})$ and this will make the total effective time interval "T" greater.

(3) From a similar argument it follows that the particles, since they undergo phase oscillations and the related radial oscillations, must now gain a greater amount of excess energy to have a radial oscillation amplitude of the same size as before. This increases the accepting phase ranges proportionately to $(\epsilon_0)^{1/2}$.

Combination of these three effects produces an increase in the effective time interval of injection proportional to ϵ_0, for small phase oscillation amplitudes.

Voltage amplitude

The variation expected in T as a result of varying the voltage amplitude applied to the accelerating electrode is shown in Figure 6. The existence of an optimum voltage has already been mentioned and is chosen so that the two limits of accepting phase range coincide. If the voltage amplitude decreases from this optimum, the stable phase ϕ_1 will increase and, the upper phase limit being $\pi - \phi_1$, the accepting half range is $\pi - 2\phi_1$ and will steadily decrease to zero. If the voltage amplitude is increased from the optimum, then ϕ_1 decreases and the upper limit, set by the radial oscillations, decreases rapidly. Then the accepting phase range goes asymptotically to zero for large voltage amplitudes.

Effective width of the accelerating chamber

By the effective width is meant the actual width minus twice the amplitude of any radial oscillations other than those already mentioned. Such oscillations arise from the circular irregularities in the magnetic field coming from eddy currents, inhomogeneities in the iron or construction. The effect of the effective width on T is illustrated in figure 6. It is seen that considerable advantage is to be gained by increasing this width.

The reason for this increase is not hard to see. A larger accelerating space means more R.F. cycles can accept particles, the number of cycles being proportional to the width of the chamber. Furthermore, the particles can now have larger radial oscillation amplitudes and this permits larger phase ranges. In the case of optimum voltage amplitudes this allows the voltage optimum to increase, thereby decreasing ϕ_1 and increasing the stable limits of phase. The increase in phase allowed on this account will involve the injection oscillations, whose amplitudes will be different for each instantaneous orbit. Thus the position of the injector will modify the effective increase in phase range gained in this way.

Depth of accelerating chamber

Besides the advantage of being able to tolerate larger vertical disturbances to the particles, increasing the accelerating chamber depth provides another advantage in the case of the Birmingham synchrotron (Oliphant, Gooden, Hide, 1947). Here it is proposed that the stable orbital plane be lowered continuously during the period of injection. The amount the orbital plane has been lowered before the particles return to the immediate neighbourhood of the injector determines the effective thickness of the proton beam. Thus a greater depth of the accelerating space allows the orbital plane to be lowered by a proportionately greater amount. The depth should consequently be as large as can be tolerated on other grounds. The depth affects very critically the energy stored in the magnetic field, other dimensions remaining constant as both the volume of space is increased and the magnetic field reduced for the same ampère-turns. It does not affect the beam-current as strongly as does the width, and so a ratio of depth to width of about 1 : 2 seems a good choice.

Table 1 illustrates what effective thicknesses of injected proton beam can be obtained for different depths of accelerating chamber for the Birmingham synchrotron.

Table 1

Depth of chamber (cm.)	7·5	10	15
Beam thickness (mm.)	2	3	4·5

Quantitative results

The general quantitative information concerning the injection process can be obtained without solving the phase equation at all. Remembering that there is a relation between the phase velocity and the difference between the orbital radius of the particle and the stable orbit radius, the analogy of the conservation of energy of the pendulum bob can be employed. The magnetic field is here assumed to rise linearly in time. A departure from this condition makes little difference to the magnitude of the results but considerably complicates the form they take. Then the maximum phase velocity and consequently the maximum orbital radius can be determined for particles entering at different phases and times.

The first relation is, in fact:

$$\frac{\partial \phi}{\partial q} = \frac{2\pi n p}{R_0}(R_0 - R) \qquad \ldots\ldots(1)$$

using the notation given in the appendix. If $\phi_0{}^u$ is the limit (upper or lower) of the acceptable phase range of the uth R.F. cycle, then, using the analogy just referred to, $\phi_0{}^u$ is given by

$$|\cos\phi_1 - \cos\phi_0{}^u - (\phi_0{}^u - \phi_1)\sin\phi_1|^{1/2}$$
$$\leqslant \frac{\sqrt{2}\pi n p (q_0 + u)^{1/2}}{R_0 A^{1/2}} \{[D - |x_1 - x_0{}^u|]^2 - (x_0{}^u)^2\}^{1/2}, \qquad \ldots\ldots(2)$$

where

$$A = \frac{V_0 c n p}{2 K R_0{}^2 (1-n)}. \qquad \ldots\ldots(3)$$

From this, by graphical methods of integration, the total effective time interval of injection T can be obtained. The number of accepting R.F. cycles is given by

$$N = \frac{2D(1-n)\epsilon_0 p}{e R_0 V_0 \sin\phi_1}. \qquad \ldots\ldots(4)$$

In the case when the phase lies on the approximately straight portion of the R.F. voltage curve, T can be integrated and becomes:

$$T = \frac{\frac{(3\cdot 3)c}{e}\sqrt{\frac{m}{e}}[n^{1/2}(1-n)^{3/2}]\left[\frac{D(D^2 - x_1{}^2)^{1/2}}{R_0{}^3}\right]}{K(V_0\cos\phi_1)^{1/2}}\epsilon_0 p^{1/2}. \qquad \ldots\ldots(5)$$

For the optimum choice of V_0, $\phi_0{}^0 = \pi - \phi_1$, and since

$$V_0 = \frac{2\pi R_0{}^2 K}{c \sin\phi_1}, \qquad \ldots\ldots(6)$$

in this case, the optimum value of ϕ_1 is given by

$$|2\cos\phi_1 - (\pi - 2\phi_1)\sin\phi_1| \operatorname{cosec}\phi_1$$
$$= \left\{\frac{np(D - x_1)}{R_0}\right\}^2 \left\{\frac{(1-n)\epsilon_0 c}{Knep}\right\}. \qquad \ldots\ldots(7)$$

For the case when $\phi - \phi_1 \ll \phi_1$ and $\phi - \phi_1 \simeq \sin(\phi - \phi_1)$, we have

$$(\pi - 2\phi_1)\left\{\tan\left(\frac{\pi - 2\phi_1}{2}\right)\right\}^{1/2} = \frac{n(D - x_1)}{R_0{}^2}\left\{\frac{2(1-n)\epsilon_0 c p}{Kne}\right\}^{1/2}, \qquad \ldots\ldots(8)$$

which determines the optimum value of ϕ_1.

§ 3. PHASE OSCILLATION AMPLITUDE BEHAVIOUR

There are no less than eight significant forces which can affect the behaviour of the phase oscillation amplitude. In order to introduce these forces in the most convenient way, those which cannot be externally varied are considered, firstly by discussing the simpler case of extreme-relativistic particles and then by extension to the case of non-relativistic particles. Intermediate regions follow immediately. There are then left four independent adjustable forces which are considered in turn. These latter can be employed to give a useful range of permissible variation of those factors which cause an increase in the phase amplitude. Finally, the effect of these various forces is considered quantitatively and curves given to enable the value of the forces to be assessed.

For an extreme-relativistic particle (i.e. sensibly constant velocity) accelerated in a synchrotron whose magnetic field increases linearly with time and whose accelerating voltage amplitude and R.F. are maintained constant, there is one damping force together with two opposing (anti-damping) forces.

The damping force can be explained in the following way. In the description of the phase and radial oscillations of section 1, it was shown that the particle reached a maximum or minimum orbital radius when the energy it received per revolution was just sufficient to maintain it on this circular orbit. This orbit marked the reversal of the phase force, as thereafter the phase acceleration changes sign, since the change in phase per revolution begins to decrease. When a particle is moving on a larger radius than that of the stable orbit, it will require a greater energy increase per revolution to maintain it on that larger radius. Thus the particle will reach its maximum or minimum radius when its phase reaches ϕ_2 (see figure 1), and it is at this point that the phase restoring force is reversed. This is the reason for the damping and the process is quite analogous to the damping of an oscillator in a viscous medium. An important observation to note here is that the rate of damping will be different for phase oscillations of different amplitudes. So long as the phases remain on the approximately straight portion of the R.F. curve, the phase velocities, and therefore the damping force, will be in proportion to the phase amplitude (cf. pendulum). Phase oscillations with amplitudes extending beyond this approximately straight portion will have their maximum phase velocities less than proportional to the amplitudes, and thus the damping rate will be less. This behaviour applies to all the forces affecting the phase amplitude.

A force which opposes this damping force is the one arising from the electric field of induction (betatron force). For the Birmingham synchrotron with a magnet yoke on the inside of the air gap, the changing return flux gives a decelerating force acting on the particle. The changing flux in the air gap gives an accelerating force, but since only part of this flux is ever enclosed by the particle orbit, the nett result is a decelerating force which decreases with increasing orbital radius. Consequently a particle increasing its orbital radius will require less energy per revolution from the R.F. accelerating field to maintain it on its radius than on the previous argument. The phase force is then reversed, not at ϕ_2, but at an intermediate value, say at ϕ_3 in figure 1. This opposing force is always a certain fraction less than unity of the damping force. The results of the above discussion are not affected by the position of the return flux from the air gap.

Another factor which increases the phase amplitude is the increase in energy of the particle. The action is not anti-damping in the sense of adding " phase energy", but is analogous to "conserving phase energy". As the energy E of the particles increases, the change in radius experienced by the particle on receiving a given excess energy of amount $\delta\epsilon$ decreases in proportion to $1/E$. But it is this change in radius per revolution which determines the "phase restoring force" and thus this restoring force is reduced ($\propto 1/E$). The frequency of the phase oscillation is reduced thereby ($\propto E^{-1/2}$) and to conserve "phase energy" the amplitude of the phase oscillation must increase ($\propto E^{1/4}$).*

These three major factors give a resultant damping to oscillations occurring on the straight portion of the R.F. curve, proportional to $E^{-1/4}$ (Veksler (1945), McMillan (1945), Bohm and Foldy (1946), Dennison and Berlin (1946), Frank (1946)). For larger amplitudes this rate of damping will decrease.

When a non-relativistic particle is accelerated, the R.F. must increase in step with the mean particle velocity. Although the same arguments as above also apply to the non-relativistic case, this changing R.F. introduces another force which still further reduces the rate of damping. Since the rate of change of R.F. is chosen to keep in step with a particle maintained on the stable orbit, it follows that the frequency will not increase quickly enough for a particle on a radius greater than the stable orbit. Hence the phase of such a particle will experience an extra acceleration, or in other words, the change in phase per revolution of this particle will increase every revolution. This means that as a particle increases its radius (decreases its phase) it also experiences an increase in its phase acceleration, or phase force, on this account. This action is just the opposite to damping and so the resultant damping due to the previous three forces is still further reduced. It is shown quantitatively below that the resultant of all these forces is to give no damping at all. Since their behaviour is dependent on the maximum phase velocity, the variations in their effects on oscillations of different amplitudes will be the same and so the damping will be zero for all oscillation amplitudes.

In the actual case of acceleration of protons to energies of about 1000 Mev., the finite relativistic effects cause a small damping action during the latter part of the acceleration.

Adjustable damping factors

Four methods exist whereby the behaviour of the phase oscillation amplitude can be adjusted. These are:—(*a*) changing the rate of change of R.F., (*b*) varying the voltage amplitude during acceleration, (*c*) varying the way in which the magnetic field increases with time, (*d*) shaping the faces of the accelerating electrode. Each of these will now be considered qualitatively and quantitatively.

Further variation in R.F.

From the argument given in the previous paragraph it follows that if the R.F. continually increases its rate of change more rapidly than to maintain a particle on any fixed radius, there will be an increase in the phase oscillation amplitude—and conversely. It also follows that this effect arising from the factor d^2v/dt^2 will depend on the change in radius of a particle for a given energy change, and this in turn will depend on the way in which the magnetic

---

* The quantitative results given here in brackets refer accurately only to small oscillation amplitudes.

field falls off radially (i.e. on n). For instance, if $n = 0$, then any finite continuous change to the R.F. will cause infinite anti-damping in the non-relativistic case, and the larger n, the less will be this effect. On this basis, it is best to design for a large n and then a large tolerance on frequency variation can be allowed.

Figure 7 shows quantitatively the continuous increase in R.F. in the case of the Birmingham synchrotron, giving a loss of about 10 % of the particles by anti-damping action. This increase in R.F. is plotted as a tolerance against the value of the magnetic field index n for different ways of increasing the magnetic field in time (different values of s).

The effect of this variable damping force is given later by equation (15). It should be noted that the rates of damping or anti-damping always decrease with increase of oscillation amplitudes, for reasons given above.

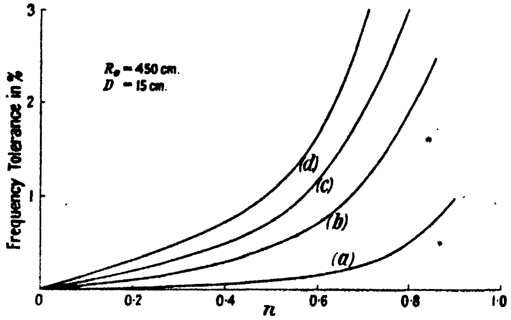


Figure 7. Variation of frequency tolerance due to anti-damping with magnetic field index n: (*a*) $s = 0{\cdot}8$, (*b*) $s = 1{\cdot}0$, (*c*) $s = 1{\cdot}2$, (*d*) $s = 1{\cdot}5$. 10 % particle loss allowed.

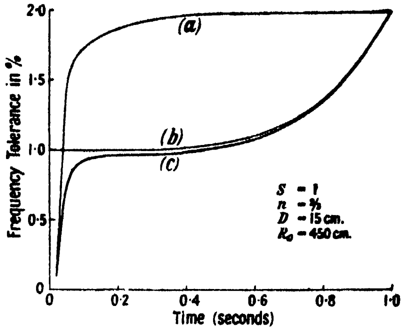


Figure 8. Variation of the frequency tolerance with time, allowing 10% loss of particles. Tolerance: (*a*) for orbital shift, (*b*) for anti-damping, (*c*) combined.

Variation of voltage amplitude

If the voltage amplitude is increased, then a particle arriving at an accelerating gap at a given phase will receive a larger excess energy, $\delta\epsilon$, than previously. In fact, $\delta\epsilon$ is proportional to V_0, the voltage amplitude, and so the rate of increase of radius is proportional to V_0. Thus by increasing the voltage amplitude during the acceleration, a particle will progressively increase its rate of increase of radius, i.e. its phase acceleration or phase restoring force ($\propto V_0$). This means that the frequency of phase oscillation is increased proportionately to $V_0^{1/2}$ and, to conserve "phase energy", the amplitude of the phase oscillation must decrease as $V_0^{-1/4}$.* Although this gives a method whereby the phase oscillations can be damped, it also means that the final width of the proton beam is larger, and this may interfere with extraction. The final width can be obtained from the relation between the radial oscillation and phase oscillation (equation 1). To obtain a reasonable effect, a large increase in voltage is required and this implies a very large increase in R.F. power.

Manner in which the magnetic field increases with time

Consider a cycle of the phase oscillations and let the magnetic field in the air gap increase as $B = Kt^s$, where t is the time and K a constant. Let $s > 1$ for

---

* See previous footnote.

the present argument; this means that the energy required to be added per revolution in order to maintain a particle in a given orbit (R_0) must increase as t^{s-1}. To maintain a particle at some constant stable phase will therefore necessitate an accelerating voltage amplitude, increasing as t^{s-1}. This increase will be finite over any one phase oscillation cycle and so the total excess energy gained by a particle undergoing the phase oscillation will be greater than in the case of constant accelerating voltage. Since the percentage increase in the magnetic field during this period is small, the particle will increase its radius more quickly than when $s=1$. This means that it reaches its maximum radius, and hence the turning point of the phase force, at a greater phase difference from ϕ_1 than previously. Hence increased damping results. The converse holds when $s<1$, which is the case of a sinusoidally increasing magnetic field.

The effect of varying s on the rate of damping is given quantitatively by equation (12). Figure 7 illustrates how varying s can be used to allow tolerable frequency variations.

Accelerating-electrode shaping

Damping is also increased if the faces of the accelerating electrodes are sloped, parallel to each other, at an angle to the radius vector, in such a way that a particle moving on a larger radius will arrive later in phase than otherwise. This will cause the particle to receive less acceleration for the same radius than without electrode shaping, an effect which, on the above arguments, will increase the damping. The quantitative value of such a system can be determined from equation (17). It is not certain how the inevitable radial disturbances thus introduced will affect the motion, but the method is an easy one with which to experiment.

Quantitative analysis

(1) *Non-relativistic case.*—Using the notation in the appendix, the phase equation for the non-relativistic particle can be written as:

$$\begin{aligned}
&\frac{\partial^2\phi}{\partial q^2}+\Big\{\frac{s}{s+1}(1-f(t))\Big[1+\frac{(\phi-\phi_1)}{2\pi pq}+\frac{1}{q}\int_0^q f(t)dq\Big]\frac{1}{q}\\
&\quad-\left(\frac{2(1-n)}{n}\right)\frac{\partial f(t)}{\partial q}\Big\}\frac{\partial\phi}{\partial q}+\frac{V_0cnp}{(s+1)KR_0^2(1-n)}\Big\{1+\frac{(\phi-\phi_1)}{2\pi pq}\\
&\quad+\frac{1}{q}\int_0^q f(t)dq+\frac{(1-n)}{2\pi np}\frac{\partial\phi}{\partial q}\Big\}\frac{\sin\phi}{q}=\frac{2\pi snp}{(s+1)(2-n)}\Big\{1+\frac{(\phi-\phi_1)}{2\pi pq}\\
&\quad+\frac{1}{q}\int_0^q f(t)dq+\frac{(1-n)}{2\pi np}.\frac{\partial\phi}{\partial q}\Big\}\Big\{1-\frac{(2-n)}{n}f(t)+\left(\frac{R_1}{R_0}\right)^{2-n}\frac{1}{(1-n)}\Big\}\frac{1}{q}\\
&\quad-2\pi p\frac{\partial f(t)}{\partial q}. \qquad \qquad \ldots\ldots(9)
\end{aligned}$$

Here the adiabatic theorem is employed, and $f(t)$ is the fractional increase in R.F. above that required to maintain the particle in the stable orbit. It is defined as

$$\text{R.F.}\equiv\nu=\frac{v_0p}{2\pi R_0}(1+f(t)). \qquad \ldots\ldots(10)$$

Neglecting $f(t)$ for the moment, the equation becomes, putting $q=x^2$,

$$\frac{\partial^2\phi}{\partial x^2}+\frac{1}{x}\frac{(s-1)}{(s+1)}\frac{\partial\phi}{\partial x}+\frac{4V_1cnp\sin\phi}{(s+1)KR_0^2(1-n)}=\frac{2\pi snp}{(s+1)(1-n)}, \qquad \ldots\ldots(11)$$

where $V_0=V_1t^{s-1}$ is the voltage amplitude.

It is readily seen that for $s=1$ the damping term disappears and the resulting equation is identical in form with that of an undamped simple pendulum with a constant torque. Thus for all amplitudes there is no damping in this case.

For small amplitudes when $\sin(\phi-\phi_1) \simeq \phi-\phi_1$ and $(\phi-\phi_1) \ll \phi_1$, the equation can be solved as a Bessel equation, the asymptotic solution being always valid. The solution is then, putting $\sin\phi_1 = \frac{2\pi s K R_0^2}{V_1 c}$, $C = \frac{V_1 c n p \cos\phi_1}{(s+1)KR_0^2(1-n)}$,

Non-relativistic:

$$\phi = \phi_1 + Gq^{\frac{1-s}{4(1+s)}} \sin\{2(Cq)^{\frac{1}{2}} + \alpha\}, \qquad \ldots\ldots(12)$$

where G and α are arbitrary constants.

(2) *Extreme-relativistic.*—For the extreme-relativistic case it can easily be shown that the solution is

$$\phi = \phi_1 + Gq^{\frac{1-2s}{4}} \sin\{2(C_1 q)^{\frac{1}{2}} + \alpha\}, \qquad \ldots\ldots(13)$$

where
$$C_1 = \frac{V_1 p c \cos\phi_1}{K(1-n)R_0^2}.$$

(3) *General case.*—In the general case intermediate between the non- and extreme-relativistic cases, and when $s=1$, the damping term can be shown to be given by

$$\text{amplitude} \propto q^{-\frac{g}{4(2+g)}}, \qquad \ldots\ldots(14)$$

where
$$g \equiv \frac{\text{kinetic energy}}{m_0 c^2}.$$

Damping effect of frequency variations follows from the general equation in the case that $(\phi-\phi_1) \ll \phi_1$ and $\sin(\phi-\phi_1) \simeq \phi-\phi_1$ by a Bessel solution. The phase damping is then found to be as q^A, where

$$A = f(t) - \frac{1}{q}\int_0^q f(t)\,dq + \frac{4(1-n)q}{n}\frac{\partial f(t)}{\partial q} - \frac{\phi-\phi_1}{2\pi p q}. \qquad \ldots\ldots(15)$$

The damping effect of voltage amplitude increase has already been shown by qualitative argument to be given by the term $V^{-1/4}$, i.e. due to this effect in the absence of others the phase varies as

$$\phi - \phi_1 = \text{const. } V^{-1/4} \cos(Dq^{1/2}V^{1/2} + \alpha), \qquad \ldots\ldots(16)$$

where D is a constant. This can be confirmed by direct solution in many cases.

The damping effect of variations of the magnetic field increase has already been given by equations (12) and (13).

The damping effect of electrode shaping is given by the relation q^{-B}, where

$$B = 1 - \frac{ps \cot\phi_1 \tan\theta}{(s+1)(1-n)}, \qquad \ldots\ldots(17)$$

where θ is the angle between the radius vector and the electrode face.

§ 4. PERMITTED VARIATIONS OF THE R.F.

Phase damping

Since the R.F. varies by a large factor, it is important to know the effects on the particle motions of small deviations of this frequency from the required law.

It has already been shown that if the frequency increases by about 2 % too rapidly, then 10 % of the particles are lost because the phase amplitude increases. Figure 7 gives the data on this effect for the Birmingham synchrotron. Figure 8 shows how in a given case this R.F. tolerance is relaxed towards the latter part of the acceleration because of the finite damping due to relativistic effects.

Orbital disturbances

It is readily seen, on the basis of the description of the particle motions given in the section on injection, that if the R.F. does not correspond to a stable orbit in the centre of the chamber, then a smaller number of particles will be accepted. In fact, the effective half width of the accelerating chamber is the shortest distance between the stable orbit and either synchrotron wall. If the R.F. changes slowly enough the particles can follow a change in the stable orbit. It then follows from the equations (1) and (11) and the reasoning in the section on injection that, in order to move the stable orbit a fraction L of the chamber half-width, $f(t)$ must obey the relation

$$f(t) \leqslant \frac{npD}{R_0}\left\{1-(1-L)\left(\frac{q}{q_0}\right)^{-1/2}\right\}. \qquad \ldots\ldots(18)$$

For this fraction L, it is safe to assume that not more than a fraction $1-(1-L)^2$ of the particles is thus lost. Figure 8 shows the tolerance permitted for the Birmingham synchrotron when 10 % of particles are lost in this way.

For rapid R.F. variations (random oscillations) the particles have difficulty in following the variations in the stable orbit and so the permitted frequency variations are increased. The general condition is given by

$$\left|-\frac{(1-n)}{2n}(\sin 2\phi_1)\frac{\partial f(t)}{\partial q}+\frac{\partial F(q)}{\partial q}+2\pi p f(t)\right| \leqslant \frac{2\pi n p D}{R_0}\left\{1-(1-L)\left(\frac{q}{q_0}\right)^{-1/2}\right\}, \qquad \ldots\ldots(19)$$

where $F(q)$ is a particular solution of the equation

$$\frac{\partial^2\phi}{\partial q^2}+\frac{s}{s+1}\frac{1}{q}\frac{\partial\phi}{\partial q}+\left\{\frac{V_1 cnp\cos\phi_1}{(s+1)KR_0^2(1-n)}\right\}\frac{\phi}{q}=-2\pi p\frac{\partial f(t)}{\partial q}.$$

Thus, for $f(t)=k\cos\omega q$, say, and for $\omega \gg \frac{1}{2q}$, it is found that $\frac{\partial F(q)}{\partial q} \simeq -2\pi p.f(t)$ and the inequality for these rapid variations becomes

$$k \leqslant \frac{2\pi n^2 p}{(1-n)(\omega\sin 2\phi_1)}\frac{D}{R}\left\{1-(1-L)\left(\frac{q}{q_0}\right)^{-1/2}\right\}. \qquad \ldots\ldots(20)$$

For the Birmingham synchrotron rapid variations in frequency can be very large indeed.

Combined effects

When these two effects of the R.F. are combined, the tolerance on the R.F., determined by allowing no more than 10 % of particles to be lost, will vary during the acceleration period. For the particular case of the Birmingham synchrotron, this variation in the tolerance is shown in figure 8. It is to be noted that the strict tolerance necessary at injection is relaxed by an order of magnitude in about 1/50 of the total acceleration time.

REFERENCES

Bohm and Foldy, 1946. *Phys. Rev.*, **70**, 249.
Dennison and Berlin, 1946. *Phys. Rev.*, **70**, 58.
Frank, 1946. *Phys. Rev.*, **70**, 174.
Kerst and Serber, 1941. *Phys. Rev.*, **60**, 53.
McMillan, 1945. *Phys. Rev.*, **68**, 143.
Oliphant, Gooden and Hide, 1947. *Proc. Phys. Soc.*, **59**, 677.
Pollock, 1946. *Phys. Rev.*, **69**, 125.
Veksler, 1945. *J. Phys. U.S.S.R.*, **9**, no. 3.

APPENDIX

Notation used :—

$\pi-\phi$ = phase of R.F. (see figure 1).
ϕ_1 = stable phase.
R = radius of particle orbit.
R_1 = max. radius of magnetic field.
$R_0 = \frac{1}{2}$ (min. radius + max. radius).
D = half width of accelerating space.
ν = radio frequency (R.F.).
$f(t)$ = relative variation in R.F. from the required law (see equation (10)).
V_0 = voltage amplitude appearing on the accelerating electrodes.
ϵ = kinetic energy.
ϵ_0 = kinetic energy at injection (time $t=t_0$).
E = total energy.
q = number of revolutions undergone by a particle starting from rest in the synchrotron, i.e. starting at zero time ($t=0$).
p = ratio of R.F. to particle frequency of revolution and is a positive integer.
T = effective time interval over which particles entering the synchrotron are accepted for acceleration to the peak energy.
t = time.
t_0 = mean time of injection.
B = flux density in the air gap and varies as $B=Kt^s\,(R_0/R)^n$, where
K = a constant, determining the rate of rise of B,
s = a positive index determining the way in which the magnetic field increases with time,
n = a positive index, less than unity, determining the way the magnetic field varies with radius.
x_1 = the radial coordinate of the injector, measured positively outwards from R_0.
$x_0{}^u+R_0$ = the radius of the instantaneous orbit of a particle entering at the uth R.F. cycle.
u = the number of R.F. cycles following the one when the instantaneous orbit has a radius R_0.
v_0 = velocity of particle in the stable orbit.

THE HOLE THEORY OF DIFFUSION

By G. WYLLIE,
Bristol

MS. received 23 *December* 1946

ABSTRACT. We show that in a dilute substitutional solid solution of one metal in another the diffusion of the solute atoms is determined by the compound activation energy $\epsilon_A + \epsilon_{ABA}$ where ϵ_A is the energy necessary for the formation of a hole next to a dissolved atom and ϵ_{ABA} the energy of activation for the hole to make one jump round that atom, provided $\epsilon_{ABA} - \epsilon_{BB} \gg kT$, where ϵ_{BB} is the energy of activation for the hole to diffuse away from the dissolved atom.

This mechanism provides an explanation of the cases where diffusion of a foreign metal atom in a lattice has a lower activation energy than self-diffusion in the same lattice, gold in lead being a conspicuous example. However, the assumption of next-neighbour interactions made in the description in this paper does not correspond to the facts of metallic structure. The statistical argument is not invalidated by this, but the calculation of the actual energies involved becomes a very difficult problem in quantum mechanics which has not yet been solved.

Johnson (1939) has given a semi-quantitative treatment of diffusion in dilute metallic solid solutions by the mechanism of Schottky defects. He pointed out that it might be energetically possible for a solute atom and a hole (i.e. a vacant lattice point) to adhere, and to wander together, with a comparatively low activation energy, through the matrix before separating. Such a process could explain the observed fact that the activation energy for diffusion of atoms dissolved substitutionally in a metal lattice is frequently less than the activation energy for self-diffusion in the same lattice.

The object of this note is to make a more complete analysis of the problem for a rather simple case, which, however, is not too far removed from experimental conditions. We consider N_A atoms of type A and N_B of type B ($N_A \ll N_B$) arranged on a cubic close-packed lattice. Then the problem is to determine the diffusion coefficient of the A atoms through the lattice, the movement of each A atom being by a jump into a neighbouring vacant lattice position. In talking about the relation of an atom to its nearest neighbours, it is useful to take a unit cell which is not face-centred but edge-centred, derived from the ordinary face-centred cell by a translation of half its edge parallel to an edge (figure 1). Let the centre atom in the cell in figure 1 be surrounded by B atoms except for the point H, which is vacant. Then there are four B atoms which lie next to the centre atom and to the vacant point, so in order to jump into the vacant position that atom has to squeeze through a gate of the form shown in figure 2, where the atoms are represented as spheres of diameter equal to the distance between the centres of nearest neighbours. Evidently it is to be expected that a foreign atom, if of small radius, should require a lower activation energy for jumping into a neighbouring hole than one of the matrix atoms (A_1 and B_1, figure 2). However, a B atom next to an

A atom, both being next to a hole, has to pass through an asymmetrical gate (A_2, figure 2), and may thus require a lower activation energy for the jump than a B atom not so situated.

We may calculate the probabilities of the different possible jumps, in terms of the energy changes involved, by the reaction rate theory developed by Eyring and others. This is done by supposing the system in an "activated state" to move in an arbitrary small length δ of the reaction coordinate about the maximum of potential energy. Then the probability of a transition per unit time is given by the product of the probability of occurrence of this activated state and the thermal velocity in the reaction coordinate, divided by δ.

If each atom is considered to vibrate in the average field due to the others, we can write down the partition function for an arbitrary configuration of the system (i.e. for arbitrary numbers of vacant lattice points and activated atoms on the way into the vacancies), taking the pressure to be constant and zero. Then the equilibrium state is given by the maximum of the partition function.

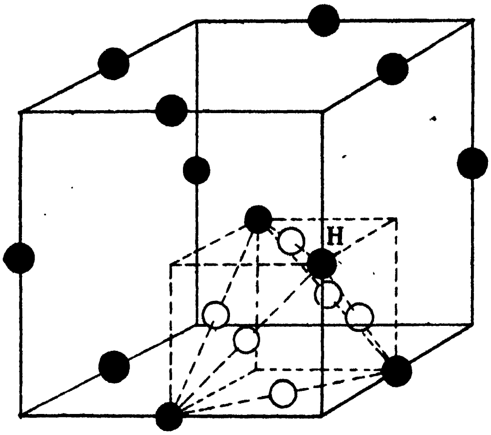


Figure 1.
Black circles indicate points of the metal lattice, white circles points of AH pair lattice.
N.B.—One black circle has been omitted for the sake of clarity.

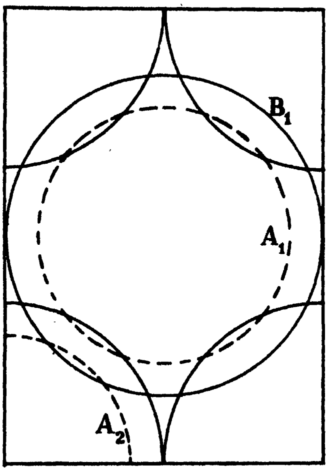


Figure 2.

Suppose n_{B} holes have only B atoms surrounding them, while n_{A} have an A atom as a nearest neighbour. We suppose also that N_{A} is so much less than N_{B} that the case of a hole having more than one A neighbour may be ignored. Of the n_{B}, let n_{BB} have an activated atom moving in. Of the n_{A}, let n_{AA} have an activated A atom moving in, n_{AB} an activated B atom not next to the A atom moving in, and n_{ABA} an activated B atom next to the A atom moving in. Let the energy required to remove a B atom from a position not next to an A atom, leaving a hole, be ϵ_{B}, and the energy to remove a B atom from a position next to an A atom be ϵ_{A}. Also let $\epsilon_{\mathrm{BB}}, \epsilon_{\mathrm{AA}}, \epsilon_{\mathrm{ABA}}, \epsilon_{\mathrm{AB}}$ ($=\epsilon_{\mathrm{BB}}$, if we consider only next-neighbour interactions) be the activation energies for the processes with the corresponding subscripts.

We neglect the changes in frequency of oscillation of atoms at the surface of holes, B atoms next an A atom and atoms between which an activated atom is just squeezing. These can be introduced if necessary in any particular case without altering the combinatory factor in the partition function, so merely modify the multiplying factors, not the exponential terms, in the final expressions for the rates. B atoms are taken to oscillate with frequency ν_B in all directions, A with ν_A in all directions, BB with ν_{BB} in both directions normal to the direction of the jump, AA with ν_{AA} in both directions, ABA with ν_{ABA1} in one direction and ν_{ABA2} in the other, AB with ν_{BB} in both. We assign arbitrary distances in the reaction coordinates $\delta_{BB, AA, ABA, AB\,(=BB)}$

The partition function F for the system is then given by

$$F = \left[\frac{N!\,(N-13N_A)!\,(12N_A)!\,12^{n_{BB}}\,.\,4^{n_{ABA}}\,.\,7^{n_{AB}}}{(N-N_A)!\,N_A!\,(12N_A-n_A)!\,n_{AA}!\,n_{ABA}!\,n_{AB}!\,n_{BB}!\times(n_A-n_{AA}-n_{ABA}-n_{AB})!\,(N-13N_A-n_B)!}\right.$$

$$\left.\times\frac{1}{(n_B-n_{BB})!}\right]\left(\frac{kT}{h\nu_B}\right)^{3(N_B-n_{BB}-n_{ABA}-n_{AB})}\cdot\left(\frac{kT}{h\nu_A}\right)^{3(N_A-n_{AA})}\cdot\left(\frac{kT}{h\nu_{BB}}\right)^{2(n_{BB}+n_{AB})}$$

$$\times\left(\frac{kT}{h\nu_{AA}}\right)^{\frac{1}{2}n_{AA}}\cdot\left(\frac{k^2T^2}{h^2\nu_{ABA1}\nu_{ABA2}}\right)^{n_{ABA}}\cdot\left(\frac{2\pi m_{BB}kT}{h^2}\delta^2_{BB}\right)^{\frac{1}{2}(n_{AB}+n_{BB})}$$

$$\times\left(\frac{2\pi m_{AA}kT}{h^2}\delta^2_{AA}\right)^{\frac{1}{2}n_{AA}}\cdot\left(\frac{2\pi m_{ABA}kT}{h^2}\delta^2_{ABA}\right)^{\frac{1}{2}n_{ABA}}$$

$$\exp\left\{-\frac{1}{kT}[E_0+n_B\epsilon_B+n_A\epsilon_A+(n_{BB}+n_{AB})\epsilon_{BB}+n_{AA}\epsilon_{AA}+n_{ABA}\epsilon_{ABA}]\right\}$$

where $N=N_B+N_A+n_B+n_A$, E_0 = potential energy of crystal with no holes, provided the temperature is sufficiently high for specific quantal effects to be neglected. This is the case for temperatures at which the interdiffusion of metals is sufficiently rapid to be of interest. m_{AA} etc. are the reduced masses for the motion in the corresponding reaction coordinates.

For the equilibrium state F, and so $\ln F$, must be a maximum for variation of n_B, n_A, n_{BB}, n_{AA}, n_{ABA}, n_{AB}. The condition that this should be so leads to the equations

$$n_{AA}=(n_A-n_{AA}-n_{AB}-n_{ABA})\frac{\nu_A^3}{\nu_{AA}^2}n_{AA}e^{-\epsilon_{AA}/kT},$$

$$n_{ABA}=(n_A-n_{AA}-n_{AB}-n_{ABA})\frac{4\nu_B^3}{\nu_{ABA1}\nu_{ABA2}}n_{ABA}e^{-\epsilon_{ABA}/kT},$$

$$n_{AB}=(n_A-n_{AA}-n_{AB}-n_{ABA})\frac{7\nu_B^3}{\nu_{BB}^2}n_{BB}e^{-\epsilon_{BB}/kT},$$

$$n_{BB}=(n_B-n_{BB})\,\frac{12\nu_B^3}{\nu_{BB}^2}\,e^{-\epsilon_{BB}/kT}\,.\,n_{BB}.$$

Also

$$n_A/n_B=y/x\,.\,e^{(\epsilon_B-\epsilon_A)/kT}\,.\,\frac{12N_A-n_A}{N_B-12N_A+n_A}=z$$

and

$$[N_B+N_A+(1+1/z)n_A]\,[N_B-12N_A+(1+1/z)n_A]=\frac{y}{z}n_A[N_B+(1+1/z)n_A]e^{\epsilon_B/kT},$$

where

$$x=\frac{n_A-n_{AA}-n_{ABA}-n_{AB}}{n_A}, \qquad y=\frac{n_B-n_{BB}}{n_B}$$

and

$$u_j=\delta_j\sqrt{\frac{2\pi m_j}{kT}}.$$

The solution of these equations is simple for $N_B \gg N_A$, $12N_A \gg n_A$, conditions corresponding to the initial physical assumptions. The velocity in the reaction coordinate j is

$$\sqrt{\frac{kT}{2\pi m_j}};$$

so we have finally for the equilibrium state $n_B=N_B e^{-\epsilon_B/kT}$, $n_A=12N_A e^{-\epsilon_A/kT}$; and the numbers of jumps of different types taking place per second are

$$n'_{BB}=n_B\frac{12\nu_B^3}{\nu_{BB}^2}e^{-\epsilon_{BB}/kT}, \qquad n'_{ABA}=n_A\frac{4\nu_B^3}{\nu_{ABA1}\nu_{ABA2}}e^{-\epsilon_{ABA}/kT},$$

$$n'_{AA}=n_A\frac{\nu_A^3}{\nu_{AA}^2}e^{-\epsilon_{AA}/kT}, \qquad n'_{AB}=n_A\frac{7\nu_B^3}{\nu_{BB}^2}e^{-\epsilon_{BB}/kT}.$$

Evidently, when a jump of the type AB takes place, the hole moves away from the A atom, whereas when a jump of type ABA takes place the hole moves round the A atom, remaining next to it. We are interested in the latter process as facilitating diffusion. Its relative probability,

$$\frac{n'_{ABA}}{n'_{ABA}+n'_{AB}}=p,$$

must then be rather large if the effect is to be important. If this is so, we must also expect n'_{AA} to be very much larger than n'_{ABA}, so that when a hole arrives next to an A atom the latter will make many jumps between the two positions available for it before any other transition takes place.

Then it appears that any of eight equally likely ABA transitions may follow, since any of the four next neighbours to the two positions concerned may jump into either position. Since there is now no point in distinguishing the position of the A atom and that of the hole, we may speak of an AH pair and denote its position by the midpoint of the line joining the two neighbouring lattice points which constitute it. Then (figure 1) the positions available for the AH pair are the centres of the faces of the cells of a cubic lattice, the edge of whose unit cube is half that of the original atomic lattice. Each point in this lattice has eight nearest neighbours, and is the centre of symmetry of those eight. By a succession of ABA transitions, the AH pair may wander through the lattice, jumping from one point to one of its nearest neighbours.

Now since every point in this lattice is a centre of symmetry for its nearest neighbours, to any jump of the AH pair corresponds an equal and opposite jump which lands it on an equivalent lattice point. Thus to any given sequence of n jumps, represented by the ordered set of vectors $(r_1, r_2, \ldots, r_k, \ldots, r_l, \ldots, r_n)$ where every r_k has magnitude r equal to half the interatomic distance in the metal lattice, corresponds the whole set of 2^n equally likely excursions ($\pm r_1, \pm r_2, \ldots, \pm r_k, \ldots, \pm r_l, \ldots, \pm r_n$). Now corresponding to any permutation of the signs of the other components, (r_k, r_l) may take the signs $(+,+)$, $(-,-)$, $(+,-)$, $(-,+)$. Thus if we take R^2, where $R = \Sigma r_k$, for each excursion, and sum over the whole set, the product terms such as $(r_k \, . \, r_l)$ cancel in groups of four and the sum is $2^n \, . \, nr^2$, so that $\overline{R^2} = nr^2$.

This gives the mean square distance travelled by an AH pair in consequence of n ABA transitions, since $\overline{R^2}$ is the same for all initial excursions $(r_1, \ldots, r_n)$ We require the mean square distance travelled by the A atom. This may initially have come from either of two positions distant r on opposite sides of the centre of the AH pair as first formed, and finally settles in either of two positions distant r on opposite sides of the final position of the centre of the AH pair. Thus, by the same argument as above, the mean square distance travelled by an A atom by the mechanism of AH pair formation, diffusion by n ABA jumps, AH pair dissociation by an AB jump, is $(n+2)r^2$. Now to a sequence of n ABA jumps preceding dissociation of the pair we must evidently assign a relative probability p^n. So the mean square distance travelled by an A atom each time a hole diffuses into a neighbouring position is

$$r^2 \frac{\sum\limits_0^\infty (n+2)p^n}{\sum\limits_0^\infty p^n} = r^2 \left[\frac{1 + \sum\limits_0^\infty (1+n)p^n}{\sum\limits_0^\infty p^n} \right]$$

$$= r^2 \, . \, \frac{2-p}{1-p}.$$

The frequency with which an AH pair is formed must equal the frequency with which a pair dissociates. The frequency of dissociation of AH pairs over the whole system is n'_{AB}, so the frequency with which a given A atom enters into an AH pair is given by

$$n'_{\mathrm{AB}}/N_{\mathrm{A}} = 12 e^{-\epsilon_{\mathrm{A}}/kT} \, . \, \frac{7\nu_{\mathrm{B}}^3}{\nu_{\mathrm{BB}}^2} e^{-\epsilon_{\mathrm{BB}}/kT}.$$

Thus the mean square distance through which an A atom diffuses in unit time by this mechanism is

$$r^2 \, . \, \frac{n'_{\mathrm{AB}}}{N_{\mathrm{A}}} \, . \, \frac{2-p}{1-p},$$

and the diffusion coefficient will be one-sixth of this; so

$$D = \frac{7}{2} a^2 \frac{\nu_{\mathrm{B}}^3}{\nu_{\mathrm{BB}}^2} e^{-(\epsilon_{\mathrm{A}}+\epsilon_{\mathrm{BB}})/kT} \, . \, \frac{\dfrac{4}{\nu_{\mathrm{ABA1}} \cdot \nu_{\mathrm{ABA2}}} e^{-\epsilon_{\mathrm{ABA}}/kT} + \dfrac{14}{\nu_{\mathrm{BB}}^2} e^{-\epsilon_{\mathrm{BB}}/kT}}{\dfrac{7}{\nu_{\mathrm{BB}}^2} e^{-\epsilon_{\mathrm{BB}}/kT}},$$

where a = interatomic distance in lattice, and so

$$D \doteqdot 2a^2 \frac{\nu_{\mathrm{B}}^3}{\nu_{\mathrm{ABA1}} \cdot \nu_{\mathrm{ABA2}}} e^{-(\epsilon_{\mathrm{A}}+\epsilon_{\mathrm{ABA}})/kT}.$$

The measured activation energy for this process, $\epsilon_A + \epsilon_{ABA}$, may be very much less than that required for self-diffusion of B atoms, which is $\epsilon_B + \epsilon_{BB}$.

It may be observed that the mean square path of an A atom for a single collision with a hole is proportional to $e^{(\epsilon_{BB} - \epsilon_{ABA})/kT}$, so should increase rapidly with decreasing temperature. This might lead to apparent anomalies in diffusion through thin metal films at relatively low temperatures in cases where the mechanism of diffusion discussed above is important.

ACKNOWLEDGMENT

In conclusion, the author wishes to express his indebtedness to Mr. F. R. N. Nabarro and especially to Dr. F. C. Frank for fruitful discussions during the preparation of this paper.

REFERENCE

JOHNSON, 1939. *Phys. Rev.*, **56**, 814.

THE IMPERIAL COLLEGE HIGH-VOLTAGE GENERATOR

BY W. B. MANN,*
National Research Council Laboratory, Chalk River, Ontario, Canada
AND
L. G. GRIMMETT, †
United Nations Educational, Scientific and Cultural Organization, Paris

MS. received 21 *October* 1946

ABSTRACT. The design and construction of two pressure-insulated electrostatic generators similar to those of Van de Graaff and Trump are briefly described. Voltage tests with one of the generators with mixtures of nitrogen and freon under pressure have shown it to be capable of producing voltages in excess of two million.

§ 1. INTRODUCTION

In the early summer of 1939 it was decided to instal a high-voltage electrostatic generator at the Imperial College to give around two million volts potential for the acceleration of positively ionized particles. The Medical Research Council had for some time also been considering a similar project for the provision of both high-voltage positive ions and electrons, for neutron, electron and x-ray investigations, but had been deterred from initiating such a programme on account of limited workshop facilities. It was therefore decided in the autumn of 1939 to make two such generators at the Imperial College to a common design following

---

* Formerly at Imperial College, London.

† Formerly at Radiotherapeutic Research Unit, Medical Research Council, Hammersmith Hospital, London.

closely that of the pressure-insulated electrostatic generators of Van de Graaff and Trump. Work on both generators was started in the workshops of the Physics department at the Imperial College towards the end of 1939. In 1941, however, the war brought this work to a standstill. At this time the high-pressure tanks had been completed and a number of parts, such as the equipotential hoops and belt-guards, the insulators, upper electrode spinning and the charging-belt pulleys, had been designed and made. Such auxiliary equipment as the compressors, belts and the 15-h.p. motor for driving the charging belt had also been delivered.

In January, 1942, the Radiotherapeutic Research Unit moved to new quarters at Hammersmith Hospital, and it was found possible to resume work on the Medical Research Council's generator, the upper and lower charging-belt pulley supports, charging-belt motor support and generating voltmeter being designed and made; the design of these latter parts was chiefly the work of Mr. J. W. Boag and Mr. P. Howard Flanders, of the Radiotherapeutic Research Unit. Work on the Imperial College generator was resumed in the early summer of 1945, the designs for the upper and lower pulley supports, driving-motor support and generating voltmeter being adopted *in toto* from the Medical Research Council generator.

Both generators have now undergone voltage tests under pressure, and it is the purpose of this paper to give the results of the tests on the I.C. generator and a description of the generators in so far as they have followed a common design. Ion-source equipment for the upper electrode is under construction, but to different designs for each generator. Load tests will therefore be the subject of future and separate publications.

§ 2. GENERATOR DESIGN

A photograph of the Imperial College generator is shown in figure 1. The hoops are formed by rolling $\frac{3}{4}$-inch diameter tubing into rings 30 inches in diameter, each separated from the next by means of three textolite insulators. Each hoop is provided with four belt-guards similar to those fitted to the Van de Graaff and Trump generators, but they are so designed that their position can be adjusted after the column has been assembled. Bakelite spacer tubes are fitted over the belt-guards on every eighth hoop up the column. A photograph of one of the hoops and three textolite insulators is shown in figure 2. The upper three dozen hoops of the I.C. generator are rhodium plated, while the same number of the M.R.C. generator hoops are nickel plated to prevent corrosion. The lower hoops are polished prior to assembling.

The column is assembled on a flat base plate in order to have a minimum possible distance between the last accelerating electrode in the discharge tube and the target. The driving motor is therefore contained in a small pressure chamber supported on the under side of the main base plate, in the manner indicated in figure 3. This figure, which was completed in 1943, shows the lower pulley support only schematically as this part had not then been designed.

The upper electrode consists of an aluminium spinning and charge is conveyed to it by means of a Tilton endless-woven cotton belt. The upper belt pulley is supported on a fixed mount while the lower belt pulley is mounted on two cantilever supports which can be adjusted so as to take up any slack in the belt. By

PROC. PHYS. SOC. VOL. 59, PT. 4 (W. B. MANN AND L. G. GRIMMETT)

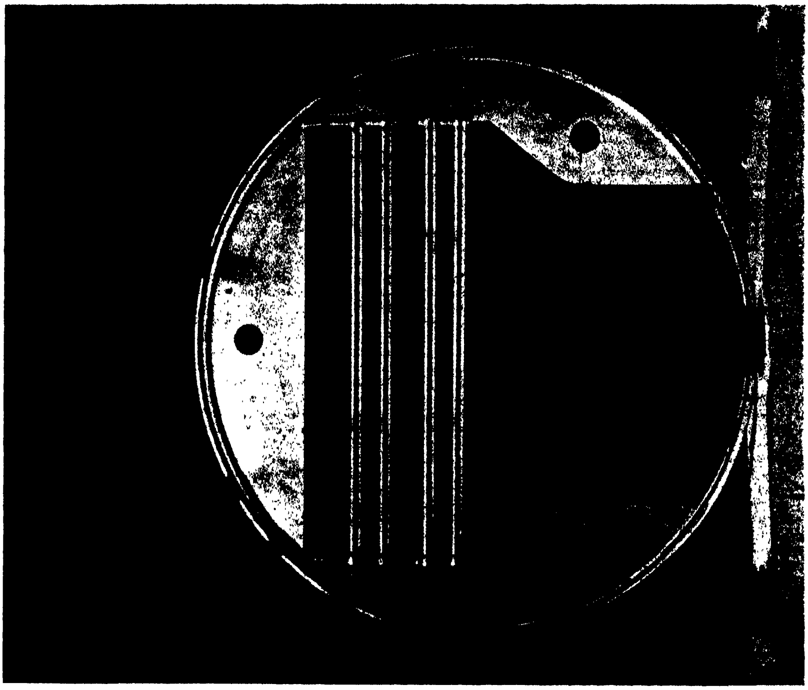


Figure 2.

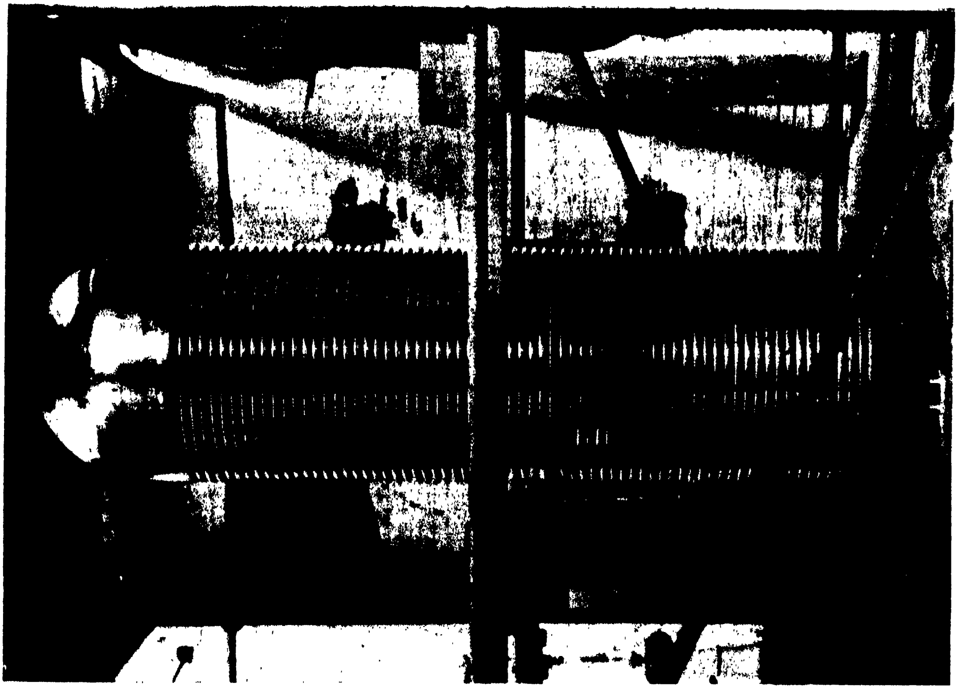


Figure 1.

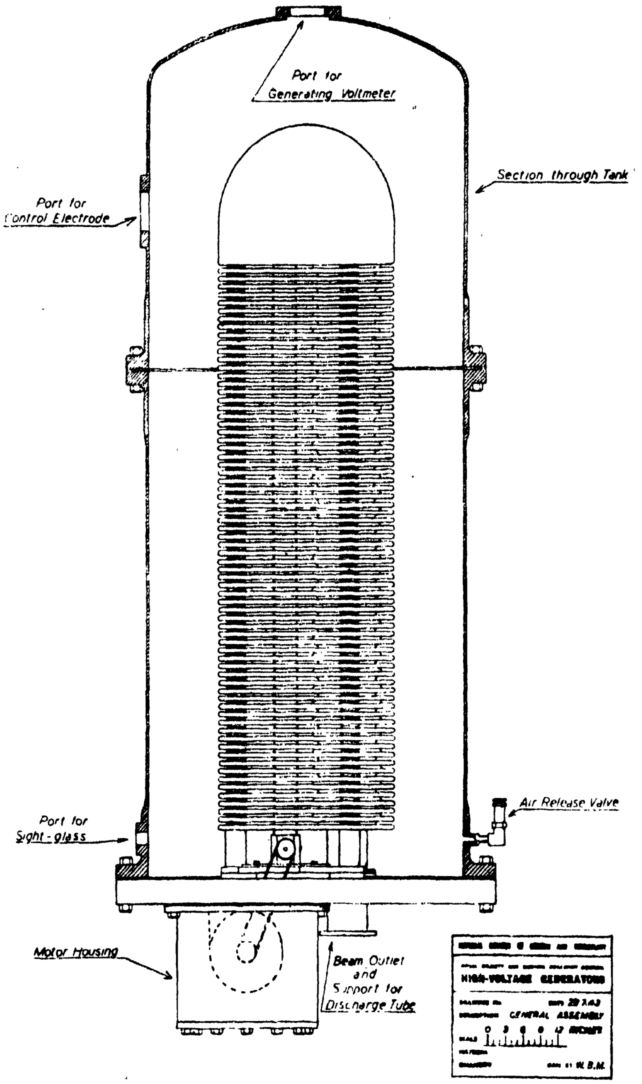
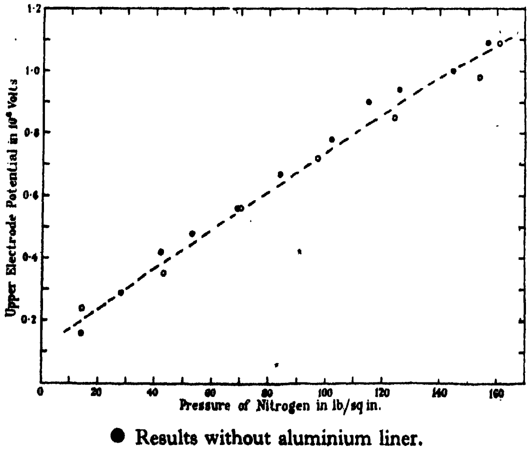
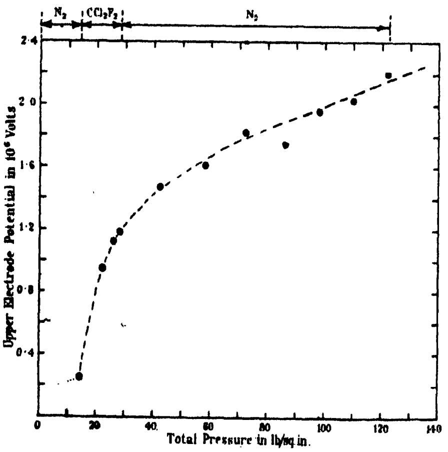

Port for Generating Voltmeter
Section through Tank
Port for Control Electrode
Air Release Valve
Port for Sight-glass
Motor Housing
Beam Outlet and Support for Discharge Tube
HIGH-VOLTAGE GENERATORS
29.7.43
GENERAL ASSEMBLY
0 3 6 9 12 INCHES
W. B. M.

Figure 3.


Upper Electrode Potential in 10^6 Volts
1·2
1·0
0·8
0·6
0·4
0·2
0
20
40
60
80
100
120
140
160
Pressure of Nitrogen in lb/sq in.
● Results without aluminium liner.


N₂
CCl₂F₂
N₂
2·4
2·0
1·6
1·2
0·8
0·4
0
20
40
60
80
100
120
140
Upper Electrode Potential in 10^6 Volts
Total Pressure in lb/sq in.

differential adjustment, on either cantilever, the belt can be made to run in a central position on the pulleys. The fulcra of these cantilever arms can be seen in the foreground of the photograph of figure 1.

The column and electrode are some 9 feet high while the Tilton endless-woven belt is 19 feet long by 20 inches wide.

The uniform distribution of potential down the column is achieved by means of a small current drain from the top electrode. This can be realized either by means of corona gaps or resistors between adjacent hoops.

The former method has been employed in the I.C. generator and the latter for the M.R.C. generator. In the case of the I.C. generator, a gap was chosen to give, with negative point to plane, around 5 μa. drain at atmospheric pressure, i.e. with the upper electrode at a positive potential of about 300 kv. Resistors of 200 megohms each between adjacent hoops were employed in the M.R.C. generator. The corona gaps were constructed simply by soldering needles to paper clips, and can easily be slipped on or off one of the plane surfaces of a hoop, these surfaces consisting of $\frac{1}{32}$-inch thick brass sheet.

§ 3. VOLTAGE TESTS

In carrying out voltage tests, the generating voltmeter was used with a balancing voltage to give zero output, the balancing voltage at the null point being observed. This method will be described in a later publication.

Aluminium liners have been provided for the upper and lower pressure tanks for both the I.C. and M.R.C. generators. The I.C. generator was tested both without and with the upper liner, but not with the lower liner in position. The results of tests in compressed nitrogen both without and with the upper liner using the I.C. generator are shown in figure 4. The results obtained for the maximum voltage as a function of pressure are not sensibly different in the two cases. The only significant difference observed was that where sparks had been fairly generally distributed over the upper electrode without the liner they tended to terminate some six inches from the lower rim of the liner when it was installed. This result might indicate that it would be better to polish the steel tank itself or to continue the upper liner a foot or two into the lower section of the steel pressure tank. The charging currents (i.e. the lower spray comb currents) required at different pressures are shown in table 1. A short-circuit test of the current delivered to the upper electrode was also made for the highest pressure without the liner, a probe being inserted through the corona-control port of the upper section of the steel tank. By means of this probe the upper electrode was short-circuited through a milliammeter to earth. On increasing the charging current to 550 μa. the short-circuit current from the upper electrode increased to 450 μa. Thereafter the short-circuit current remained constant at 450 μa. with increase of the charging current to 1·5 ma. The theoretical saturation current for such a belt should be in the neighbourhood of 1·5 ma., and the discrepancy may be accounted for by leakage through the belt. It is to be hoped, therefore, that there will be an improvement in the current available for use in the discharge tube as the belt dries with use. The current loss due to corona is, as indicated by the figures of table 1, not large.

The voltage obtained at 1 atmosphere with nitrogen was observed always to be lower than that obtained with air at the same pressure. This is in agreement with

Table 1

Voltage test without liner			Voltage test with liner		
Pressure of nitrogen (lb./sq. in.)	Spray comb current (microamp.)	Voltage	Pressure of nitrogen (lb./sq. in.)	Spray comb current (microamp.)	Voltage
14	—	$0·16\times10^{6}$	14	40	$0·24\times10^{6}$
28	—	0·29	43	50	0·35
42	—	0·42	70	75	0·56
53	—	0·48	97	90	0·72
69	—	0·56	124	100	0·85
84	—	0·67	154	115	0·98
102	—	0·78	161	120	1·09
115		0·90			
126		0·94			
145		1·00			
157		1·09			

the fact that the dielectric strength of air is greater than that of nitrogen on account of the presence of negative oxygen ions.

Finally, measurements were made with the I.C. generator using mixtures of nitrogen and freon to give increased dielectric strength. The results obtained are shown in table 2 and figure 5.

Table 2

Pressure of nitrogen (lb./sq. in.)	Pressure of freon (lb./sq. in.)	Spray comb current (microamp.)	Voltage
14	8	100	$0·95\times10^{6}$
14	12	120	1·12
14	14	100	1·18
27·5	14	120	1·47
44	14	140	1·61
58	14	150	1·82
72	14	100	1·75
84	14	120	1·96
96	14	120	2·03
108	14	160	2·20

The upper electrode, which was only resting in position, appeared to become unstable at the highest values of voltage and an intense rumbling noise developed. With the upper electrode securely fastened to the column, however, there appears to be no reason why the (voltage, pressure) curve should not be continued to around $2·5\times10^{6}$ volts.

§ 4. CONCLUSION

At the time of writing, the ion source, accelerating tube and vacuum pumping systems are not yet finished. The voltage supply for the ion source inside the upper electrode is now complete but is, as mentioned previously, of a different design from that of the M.R.C. generator. Results for current output will therefore be the subject of future, separate, publications; this present paper deals only with those features of design which are common to both generators. Of the present authors, one (W. B. M.) has been associated with the project since its inception in 1939 till the present time, with an absence of some four years on war work; the other (L. G. G.) was associated with the work from its inception in 1939 until the autumn of 1944, besides being actively concerned for some time prior to 1939 with the question of providing such a high-voltage source for medica. work. In addition, Mr. J. W. Boag and Mr. P. Howard Flanders, of the Radiotherapeutic Research Unit, have been associated with the work on the M.R.C. generator since 1942.

§ 5. ACKNOWLEDGMENTS

Much of the work on the generators has been done in the Physics department workshop of the Imperial College, all but the heavy engineering parts of the I.C. generator having been made there and also the hoops and belt-guards, upper and lower pulleys and insulators of the M.R.C. generator. We are indebted to the following for their assistance and excellent work: Mr. W. Thompson, Mr. A. E. Davis, Mr. A. H. Bridger, Mr. F. Hart, Mr. L. Jenkins, Mr. S. Leach, Mr. W. Shand and Mr. W. S. Tonks. We are also indebted to Mr. C. W. Knighton and Mr. A. J. Campbell, of Messrs. Babcock and Wilcox Ltd., for the kind personal interest they took in the design and construction of the pressure tanks.

In the erection and testing of the I.C. generator one of us (W. B. M.) wishes gratefully to acknowledge the assistance which he has received from Dr. J. L. O. G. Michiels and the following research students and visitors: Mr. H. A. Dell, Mr. G. W. Green, Mr. G. A. Mann, Dr. P. Mottier, Mr. P. L. Parsons and Mr. G. C. Tavernier. Grateful acknowledgment is also made of the kind assistance and advice received from Mr. E. H. Laister, of Messrs. Johnson, Matthey and Co. Ltd., in connection with the rhodium plating of the brass hoops and from Mr. R. H. Walter and Mr. A. C. Whitney, of the Air Ministry, in connection with the installation of the high-pressure panel for handling nitrogen which was supplied by the Air Ministry in cylinders at 250 atmospheres. He also wishes to acknowledge with thanks a grant received from the Warren Fund of the Royal Society to cover the expenses of a trip to the United States in 1939. The cost of the I.C. generator has also been met by grants made from this fund.

The other author (L. G. G.) wishes to acknowledge with thanks the assistance rendered by Mr. J. W. Boag, Mr. P. Howard-Flanders, Mr. E. A. Rendle, and Mr. F. D. Pilling, of the Radiotherapeutic Research Unit.

We are both indebted to Professor Sir George Thomson for his interest in the work since its inception.

ON THE DETERMINATION OF ASPHERIC PROFILES

By E. WOLF and W. S. PREDDY,
University of Bristol

MS. received 1 *November* 1946

ABSTRACT. Exact parametric equations are deduced for the profiles of plano-aspheric lenses designed to produce axial stigmatism in a given axially symmetric pencil of rays. These formulae take an agreeably simple form in the special case where the point of stigmatism is at infinity. As an application, parametric equations are deduced for the corrector plate of the Schmidt Camera.

§ 1. INTRODUCTION

In many optical systems involving an aspheric surface, this surface, whose essential function is to improve the performance over the whole of a finite field, is designed so as to bring accurately to a focus the rays proceeding from the axial point of a selected object surface, which may be at infinity. Various exact and approximate formulae for the shape of such surfaces have been given in particular cases. In other cases, methods of successive approximation have been used, since a straightforward application of Snell's law leads to differential equations which are inconvenient to work with.

Using the principle of equal optical path, we deduce, with the help of a result proved in the next section, exact formulae for the surface-profile needed to annul the zonal aberrations of a given wave front. These formulae take an agreeably simple form in the special case where the point of stigmatism is at infinity. Although these formulae are exact, they involve an integral which in general has to be evaluated numerically.

As an application, parametric equations for the plate profile in the classical Schmidt Camera are obtained which have a wider range of validity than the formulae given by B. Strömgren (1935) and by J. G. Baker (1940).

§ 2. A PRELIMINARY RESULT

Let OH be the ordinate at a point O on the axis of symmetry of a system of rays proceeding from an axial source. We first derive an expression for the optical path difference between any two rays before reaching OH.

We denote by W a wave front corresponding to the system. Any surface which is orthogonal to all the rays is such a wave front, and the optical distance (O.D.) from the source to W is the same for every ray. Consider two neighbouring rays meeting W in A and C, and OH in B and D at heights h and $h+\delta h$ respectively. Let ω_1 denote the angle which the ray AB makes with the axis. Through B draw a line perpendicular to AB meeting CD in E. Finally let AB be denoted by I_h and CD by $I_{h+\delta h}$ (figure 1).

Then BE is a tangent to the wave surface passing through B, whence

$$I_{h+\delta h}-I_h=-\delta h\sin\omega_1+O((\delta h)^2).$$

Dividing by δh, and proceeding to the limit as $\delta h\to 0$, we obtain

$$\frac{dI_h}{dh}=-\sin\omega_1,$$

whence

$$I_h=-\int^h\sin\omega_1\,dh.$$

Let $I_{h_1}^{h_2}$ be the optical path difference between two rays which reach OH at heights h_2 and h_1.

Then

$$I_{h_1}^{h_2}=I_{h_2}-I_{h_1}=-\int_{h_1}^{h_2}\sin\omega_1\,dh. \qquad \ldots\ldots(2.1)$$

§ 3. PLANO-ASPHERIC LENSES

With the help of (2.1) we can determine the profile of the aspheric lens which will eliminate the zonal aberrations of a given incoming wave front. Only the case of aspheric lenses with one face plane is considered here, but it will be seen that the methods of §§ 3.1 and 4.1 (amounting essentially to the use of an ikonal) apply without change to the more general case where one face of the lens is of a given profile and the other face (the profile of which is to be determined to give axial stigmatism) is the last refracting surface of the system.*

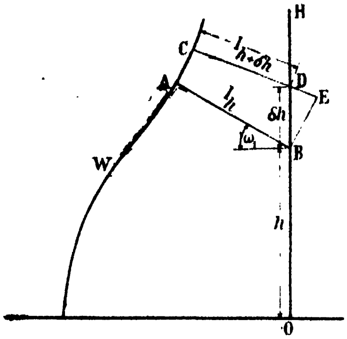


Figure 1.

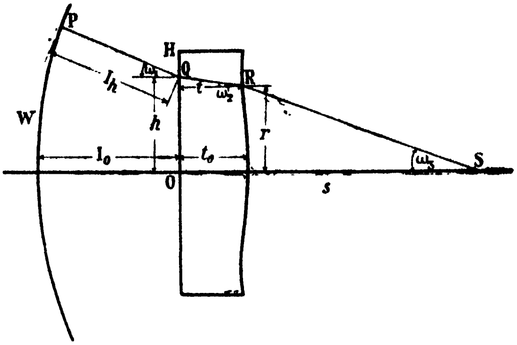


Figure 2.

§ 3.1. We first examine the mathematically simpler case where the aspheric surface of the lens faces the point of stigmatism. Figure 2 shows a ray, PQRS, proceeding from the wave front W towards the point of stigmatism S and meeting the refracting surfaces in Q and R. The origin O is taken at the point where the axis intersects the plane face of the lens.

D_h, the O.D. between P and S, has to be the same for every ray. We have

$$D_h=I_h+\frac{nt}{\cos\omega_2}+\frac{s-t}{\cos\omega_3},$$

and for the axial ray

$$D_0=I_0+nt_0+s-t_0,$$

where t and t_0 are the thicknesses of the lens at R and O, s denotes† the distance

---

* This case has been treated also on the basis of the ikonal theory by Luneberg (1944). A different method, based on successive approximation, is described by Herzberger and Hoadley (1946); it is applicable also to the calculation of aspheric surfaces in the interior of a system.

† The case of a virtual image is also covered if s is allowed to take negative values.

OS, ω_1, ω_2 and ω_3 are the angles which PQ, QR, RS make with the axis (see figure 2), and n is the refractive index of the material of the plate.

Equating D_h to D_0, we obtain

$$I_0^h+\frac{nt}{\cos\omega_2}+\frac{s-t}{\cos\omega_3}-(n-1)t_0-s=0. \qquad \ldots\ldots(3.11)$$

From the figure

$$\cos\omega_3=\frac{s-t}{[(s-t)^2+(h-t\tan\omega_2)^2]^{1/2}}. \qquad \ldots\ldots(3.12)$$

Substituting for $\cos\omega_3$ in (3.11) we obtain

$$I_0^h+\frac{nt}{\cos\omega_2}+[(s-t)^2+(h-t\tan\omega_2)^2]^{1/2}-(n-1)t_0-s=0, \qquad \ldots\ldots(3.13)$$

where

$$\omega_2=\sin^{-1}\left(\frac{1}{n}\sin\omega\right). \qquad \ldots\ldots(3.14)$$

Rationalizing and rearranging (3.13) and substituting for I_0^h in terms of ω_2,* we find that t satisfies the quadratic equation

$$\left.\begin{aligned}&At^2+2Bt\cos\omega_2+C\cos^2\omega_2=0,\\ &\text{where}\\ &A=n^2-1,\\ &B=h\sin\omega_2+s\cos\omega_2-n[s+(n-1)t_0+n\int_0^h\sin\omega_2\,dh]\\ &C=[2s+(n-1)t_0+n\int_0^h\sin\omega_2 dh][(n-1)t_0+n\int_0^h\sin\omega_2 dh]-h^2.\end{aligned}\right\} \qquad \ldots\ldots(3.15)$$

The roots of this equation are

$$t_1=\frac{\cos\omega_2}{A}[-B+\Delta^{1/2}],\quad t_2=\frac{\cos\omega_2}{A}[-B-\Delta^{1/2}], \qquad \ldots\ldots(3.16)$$

where

$$\Delta=B^2-AC. \qquad \ldots\ldots(3.17)$$

It can easily be shown that only the solution $t=t_2$ satisfies the physical conditions.

From the figure, the radius r corresponding to h is given by

$$r=h-t\tan\omega_2, \qquad \ldots\ldots(3.18)$$

so that finally we have

$$t+ir=\frac{-e^{-i\omega_2}}{n^2-1}[B+\Delta^{1/2}]+ih. \qquad \ldots\ldots(3.19)$$

This equation gives the exact profile of the lens in terms of the parameter h.

§3.2. The case when the plane face of the lens is nearer to the point of stigmatism can be dealt with in a similar manner. Let O now be the point of intersection of the axis with the aspheric face of the lens and let t denote the thickness of the lens at Q (figure 3).

Then

$$D_h=I_h+\frac{t_0-t}{\cos\omega_1}+\frac{nt}{\cos\omega_2}+\frac{s-t_0}{\cos\omega_3}$$

and

$$D_0=I_0+nt_0+s-t_0.$$

---

* Not ω_1, as might appear more natural; the final solution takes a simpler form when expressed in terms of ω_2.

Equating D_h to D_0, we obtain

$$\frac{t_0-t}{\cos\omega_1}+\frac{nt}{\cos\omega_2}+\frac{s-t_0}{\cos\omega_3}-(n-1)t_0-s-\int_0^h \sin\omega_1\,dh=0 \qquad \ldots\ldots(3.21)$$

From the figure

$$\tan\omega_3=\frac{h-(t_0-t)\tan\omega_1-t\tan\omega_2}{s-t_0}\,; \qquad \ldots\ldots(3.22)$$

also

$$n\sin\omega_2=\sin\omega_3. \qquad \ldots\ldots(3.23)$$

ω_2 and ω_3 can be eliminated between (3.21), (3.22) and (3.23), to give the equation for t; it is easier, however, substitute for ω_3 in terms of ω_2, insert values for the known quantities and solve the resulting equations by numerical methods.

The corresponding radius r is given by

$$r=h-(t_0-t)\tan\omega_1. \qquad \ldots\ldots(3.24)$$

§ 4. PLANO-ASPHERIC LENSES: OBJECT OR IMAGE AT INFINITY

In the limiting case when $s\to\infty$, the lens converts into each other two ray-systems of which one is a parallel beam.

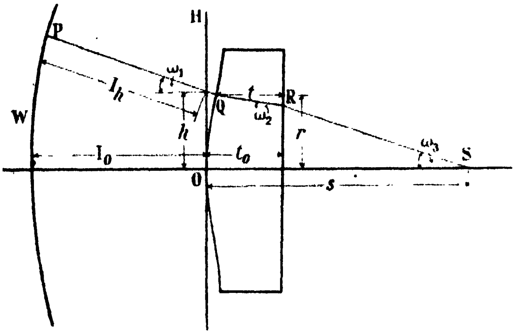


Figure 3.

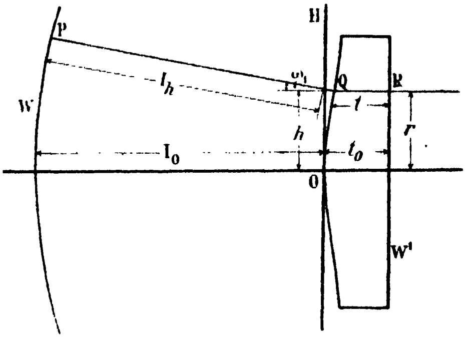


Figure 4.

§ 4.1. First we consider the case when the lens has its aspheric side towards the parallel beam. Dividing (3.15) by s and proceeding to the limit as $s\to\infty$, we obtain

$$2t\cos\omega_2(\cos\omega_2-n)+2\cos^2\omega_2[(n-1)t_0-I_0^h]=0$$

or

$$t=\frac{\cos\omega_2[(n-1)t_0-I_0^h]}{n-\cos\omega_2}, \qquad \ldots\ldots(4.11)$$

and, as in § 3.1,

$$r=h-t\tan\omega_2. \qquad \ldots\ldots(4.12)$$

On substituting for I_0^h the complete solution can finally be expressed in the form

$$t+ir=\frac{e^{-i\omega_2}}{n-\cos\omega_2}[(n-1)t_0+n\int_0^h \sin\omega_2\,dh]+ih, \qquad \ldots\ldots(4.13)$$

where

$$\omega_2=\sin^{-1}\left(\frac{1}{n}\sin\omega_1\right). \qquad \ldots\ldots(4.14)$$

§ 4.2. The solution for the case where the lens has its plane side towards the parallel beam will now be obtained directly.

Since all the rays meet the plane face at right angles, it is a wave front (W′) of the system (figure 4).

The O.D. between the wave fronts W and W′ for the general ray is

$$D_h = I_h + \frac{t_0 - t}{\cos \omega_1} + nt$$

and for the axial ray is

$$D_0 = I_0 + nt_0,$$

whence

$$I_0^h + \frac{t_0 - t}{\cos \omega_1} + nt - nt_0 = 0,$$

giving

$$t = t_0 - \frac{I_0^h \cos \omega_1}{n \cos \omega_1 - 1}. \qquad \ldots\ldots(4.21)$$

We also have

$$r = h - (t_0 - t) \tan \omega_1. \qquad \ldots\ldots(4.22)$$

Substituting for I_0^h we finally obtain

$$t + ir = t_0 + ih + \frac{e^{i\omega_1}}{n \cos \omega_1 - 1} \int_0^h \sin \omega_1 \, dh. \qquad \ldots\ldots(4.23)$$

§ 5. METHODS OF CALCULATION

In systems for which the explicit relation between ω_1 and h cannot be easily obtained, it is usually more practical to evaluate $\int_0^h \sin \omega_1 \, dh$ by numerical integration from a ray trace, or from power-series expansions for $\sin \omega_1$.

By the application of the formulae deduced in this paper, the thickness of the aspheric lens and its corresponding radius can then be calculated for every value of the parameter h for which ω_1 and $\int_0^h \sin \omega_1 \, dh$ have been determined. If more values are required they may be obtained by interpolation or curve fitting.

§ 6. APPLICATION: THE SCHMIDT CAMERA

We now apply equations (4.13) and (4.23) to find the profile of the figured surface of the Schmidt Camera. This system consists of a sperical mirror M and an aspheric plate P situated at its centre of curvature C. We first examine the arrangement (employed by Schmidt himself) in which the plane side of the plate faces the mirror (figure 5).

The rays which enter in a direction parallel to the axis pass through P, are reflected by M and focus at a point F near the paraxial focus of M. The figuring on the plate is such that it eliminates the axial spherical aberration of the system. We take the radius of the mirror as unity and denote CF by f. Further, we denote by δ the distance CO (measured as positive towards the mirror), O being, as in § 4.1, the point of intersection of the plane face with the axis. Finally we take as parameter of the ray-system the angle ϕ between the axis and the line joining C with the point where the rays reach M.

From the geometry of the figure we find that

$$\omega_1=\tan^{-1}\frac{\sin\phi-f\sin 2\phi}{\cos\phi-f\cos 2\phi}, \qquad \ldots\ldots(6.1)$$

$$h=\frac{\sin\phi(f+\delta)-f\delta\sin 2\phi}{\cos\phi-f\cos 2\phi}, \qquad \ldots\ldots(6.2)$$

$$I_0^h=\frac{2\cos\phi-f\cos 2\phi-\delta}{\cos\phi-f\cos 2\phi}\cdot(1-2f\cos\phi+f^2)^{1/2}-(2-\delta-f). \quad \ldots\ldots(6.3)$$

In substituting into (4.13) in terms of ϕ and equating real and imaginary parts, we obtain the following exact parametric equations for the plate profile

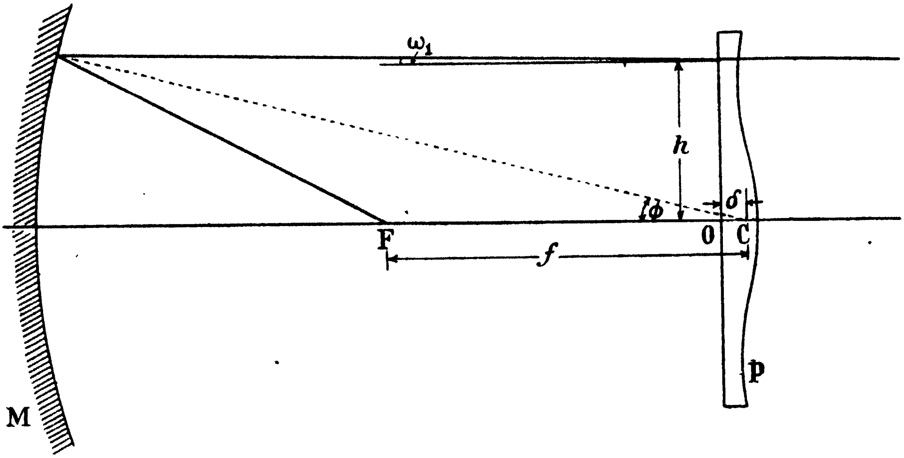


Figure 5.

if, as before, t denotes the thickness of the plate and r its corresponding zonal radius:

$$t=\frac{[n^2(1-2f\cos\phi+f^2)-\sin^2\phi(1-2f\cos\phi)^2]^{1/2}[(n-1)t_0-I_0^h]}{n^2(1-2f\cos\phi+f^2)^{1/2}-[n^2(1-2f\cos\phi+f^2)-\sin^2\phi(1-2f\cos\phi)^2]^{1/2}} \qquad \ldots\ldots(6.4)$$

and

$$r=\sin\phi\left\{\frac{f+\delta-2f\delta\cos\phi}{\cos\phi-f\cos 2\phi}-\frac{(1-2f\cos\phi)[(n-1)t_0-I_0^h]}{n^2(1-2f\cos\phi+f^2)^{1/2}-[n^2(1-2f\cos\phi+f^2)-\sin^2\phi(1-2f\cos\phi)^2]^{1/2}}\right\}, \qquad \ldots\ldots(6.5)$$

where I_0^h is given by (6.3).

The focal length CF $=f$ is usually chosen so that the chromatic aberration introduced by the plate is minimized. There are several different ways in which minimum chromatic aberration can be defined. B. Strömgren (1935) has shown that if h_n represents the height of the neutral zone (i.e. the radius of the zone for which a ray parallel to the axis passes through the plate undeviated) * and h_a the aperture radius, then to minimize (in Seidel approximation) the greatest angular departure from flatness over the whole plate, f should be chosen

* In terms of h_n, $f=\frac{1}{2}(1-h_n^{2})^{1/2}$, as can be easily deduced from the figure.

so that $h_n/h_a = 0{\cdot}866$. With this choice of f the maximum deviation of the ray parallel to the axis in the convergent sense is (again neglecting higher-order terms) the same as the maximum deviation in the divergent sense. Lucy (1940) showed that we minimize the integral of all deviations over the whole area of the aperture by giving the ratio h_n/h_a a value approximately equal to 0·79.

To eliminate primary coma, the distance δ from C to the point O where the plane face meets the axis should be approximately equal * to t_0/n.

The camera may also be designed with the aspheric surface of the plate facing the mirror.† In this arrangement the aspheric surface passes through C (figure 6) so that ω, h and I_0^h are given by (6.1), (6.2) and (6.3) with $\delta = 0$. Then on substituting in (4.23) in terms of ϕ and equating real and imaginary parts we obtain the following equations for the plate profile, equivalent to (8) and (10) of Lucy's paper (1941):

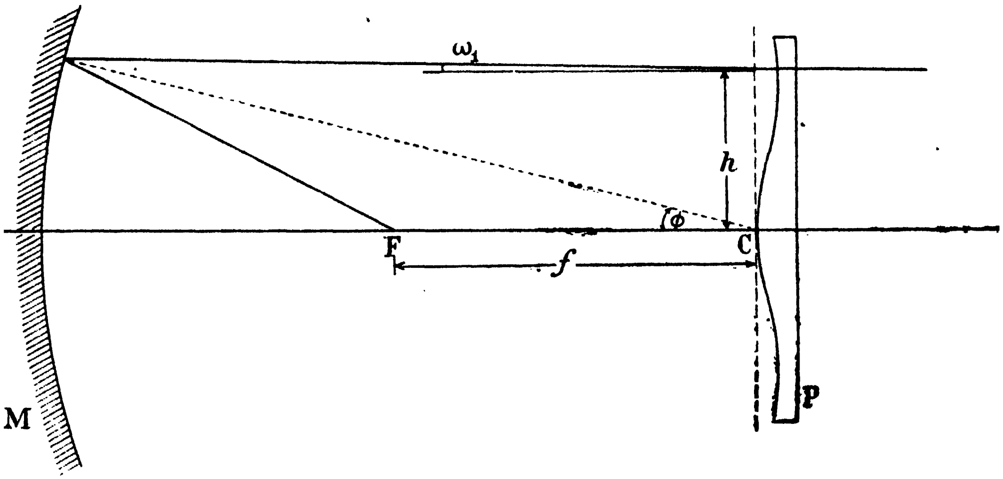


Figure 6.

$$t = t_0 - \frac{(2\cos\phi - f\cos 2\phi)(1-2f\cos\phi+f^2)^{1/2} - (2-f)(\cos\phi - f\cos 2\phi)}{n(\cos\phi - f\cos 2\phi) - (1-2f\cos\phi+f^2)^{1/2}} \quad \ldots\ldots(6.6)$$

and

$$r = \frac{\sin\phi}{\cos\phi - f\cos 2\phi}\left[f - (1-2f\cos\phi). \frac{(2\cos\phi - f\cos 2\phi)(1-2f\cos\phi+f^2)^{1/2} - (2-f)(\cos\phi - f\cos 2\phi)}{n(\cos\phi - f\cos 2\phi) - (1-2f\cos\phi+f^2)^{1/2}}\right]. \quad \ldots\ldots(6.7)$$

For most astronomical purposes, the power-series expansions for the Schmidt Camera given by B. Strömgren (1935) and extended by J. G. Baker (1940) are sufficiently accurate and are simpler to work with than the formulae of this section. For sufficiently wide-aperture systems, however, these formulae are no longer

---

* In Lucy's paper the plate is incorrectly placed with its plane face passing through C (i.e. $\delta = 0$). Equations (6.4) and (6.5) become equivalent to Lucy's results if δ is given this value.

† A procedure apparently adopted by some authors for the sake of mathematical convenience. In the astronomical case it has the effect of producing sharp "ghosts" unless the plate is given a slight overall "bending". Such a bending has little effect on the image errors, but it complicates the practical construction of the plate.

adequate. As J. G. Baker (1940) remarks: "for such extreme cases either a differential correction or else an integration based on ray-tracing formulae must be carried through". In the present section we have provided formulae suited to such extreme cases.

ACKNOWLEDGMENT

In conclusion we should like to express our thanks to Dr. E. H. Linfoot for much valuable assistance with the preparation of this paper.

REFERENCES

Baker, J. G., 1940. *Proc. Amer. Phil. Soc.*, **82**, 323.
Carathéodory, C., 1940. *Hamburg. Math. Einzelschr.*, **28**.
Glancy, A. E., 1946. *J. Opt. Soc. Amer.*, **36**, 416.
Herzberger, M. and Hoadley, H. O., 1946. *J. Opt. Soc. Amer.*, **36**, 334.
Lucy, F. A., 1940. *J. Opt. Soc. Amer.*, **30**, 251 ; 1941. *Ibid.*, **31**, 358.
Luneberg, R. K., 1944. *Mathematical Theory of Optics* (Brown University).
Strömgren, B., 1935. *Vierteljahreschrift der Astronom. Ges.*, **70**, 65.

CORRIGENDA

"The fundamental concepts concerning surface tension and capillarity", by R. C. Brown (*Proc. Phys. Soc.*, **59**, 429 (1947)).

Page 436, equation (4): insert minus sign in front of "$p_s t$".
Page 445, equation (12): for "$-\gamma_{S(S)}$" read "$-\gamma_{L(S)}$".
Page 445, beginning of seventh line below the figures: for "$-\gamma_{L(L)}$" read "$\gamma_{L(S)}$".

"The short-period time-variation of the luminescence of a zinc sulphide phosphor under ultra-violet excitation", by Mary P. Lord, A. L. G. Rees and M. E. Wise (*Proc. Phys. Soc.*, **59**, 473 (1947)).

Page 473, insert "Material Research Laboratory" before "Philips' Lamps Ltd".
Page 477, line 3, insert after "present". "The phosphor also contained 0·48% magnesium, which does not cause activation." (The authors thank Mr. C. G. A. Hill for this information.)

REVIEWS OF BOOKS

Theory and Application of Mathieu Functions, by N. W. McLachlan. Pp. xii + 401. (London: Geoffrey Cumberlege, at the Oxford University Press, 1947.) 42*s*.

Mathieu functions satisfy the differential equation

$$\frac{d^2y}{dz^2}+(a-2q\cos 2z)y=0, \qquad \ldots\ldots(1)$$

just as Bessel functions satisfy

$$\frac{d^2y}{dz^2}+\frac{1}{z}\frac{dy}{dz}+\left(1-\frac{n^2}{z^2}\right)y=0.$$

In both instances the equation defines y as a function of z and of certain parameters (a and q for the former, n for the latter), and we have to include the case where z is pure imaginary. For both equations we then write the solution as a function of z/i and call it a *modified* function.

In the case of Bessel equation, there can only be two independent solutions, and the general solution is a linear combination of these. Yet we are familiar with the fact that particular combinations are of such frequent occurrence, and such general usefulness, that we treat them as independent solutions of equal standing, so that we have not only J_n and Y_n, but also *ber* and *bei*, *ker* and *kei*, $H_n^{(1)}$ and $H_n^{(2)}$, K_n and N_n as well as I_n, and we expect a reader to be familiar with all of them.

The Bessel functions are much better known than those of Mathieu, both in the sense that many people are familiar with the properties of them, and also in the sense that more of their properties have been sought out and placed on record in mathematical literature. This arises from a combination of many causes, of which the greater intrinsic complexity of the study of Mathieu functions, and the fact that applied mathematicians had not called urgently for them, are doubtless important. The instrinsic difficulty arises in part from the fact that there are two parameters, and in part from the fact that there is a trigonometric instead of an algebraic term in the differential equation. Applied mathematicians did not ask for them, they would say, because their investigations did not lead that way; but we may suspect that they tended to avoid investigation which would have called for Mathieu functions, just because their properties were not fully catalogued and particularly because there were no tables of them—a fact which in its turn may be traced back, at least in part, to the existence of the multiplicity of parameters.

Nevertheless, there is a literature, and Dr. McLachlan has performed a really stupendous task in sorting out and presenting afresh, and in a co-ordinated manner, the whole theory of the functions. Naturally, he has had to fill in many of the gaps, and has done so with a modesty which makes it difficult to pick out for mention his own original contributions.

As with the Bessel functions, it is found desirable to define and work with a number of different fundamental pairs of solutions, and these are here systematized under a notation which makes it fairly easy to keep their special peculiarities in mind. These functions, however, possess one property which has no analogue in Bessel functions, though it has analogies in other equations familiar to the physicist. If we seek a solution of (1) which shall be periodic in z (as occurs naturally, if z is an angle), we find that there must be a relation between a and q. In other words, there are then *characteristic values* of a, imposed by the very nature of the solution. This is a situation familiar in wave mechanics, where the requirement that a solution of Schrödinger's equation shall be regular at infinity imposes characteristic values on the energy. Dr. McLachlan gives separate consideration to solutions of this periodic type, and puts workers much more deeply in his debt by considering in adequate detail the problem of numerical calculation of the solutions and of the characteristic numbers.

Two of the most useful types of solution appear to be those obtained by expansion in terms of trigonometric functions on the one hand and in terms of Bessel functions on the other. The general properties of orthogonality, so important for fitting solutions to boundary conditions, are treated, and asymptotic expansions are fully dealt with. A most valuable appendix, due to the late Prof. Ince, gives a table of the characteristic numbers. Ince, in fact, did a great deal of work on these functions, and published relatively extensive tables which are not reproduced in this book, though references are supplied.

Not only has Dr. McLachlan given us the whole corpus of useful recorded knowledge of these functions, but he discusses also the physical problems in which they play a prominent part. One such problem arises when the wave equation in elliptic co-ordinates is *separated*. For the one variable, the result is an ordinary Mathieu equation, and for the other a *modified* one, and it was in this connexion that Mathieu was first led to study the functions. Lunar theory leads to a differential equation which is a generalization of Mathieu's, but the theory of the two is very similar, and MacLachlan includes a chapter on the subject. Other physical applications have mostly arisen in recent years—Oseen's hydrodynamic equation, the theory of frequency modulation and of wave guides—but a striking exception is the detailed explanation of Melde's experiment with a tuning fork, where the string maintained in oscillation by it has a frequency which is a sub-multple of that of the fork.

It is clear that the contemporary development of physics will cause a demand for these functions. It is at least probable that this book will encourage the examination of phenomena in which these functions are involved, and which, without it, would have been set aside for future examination.

J. H. A.

THE PROCEEDINGS OF THE PHYSICAL SOCIETY

VOL. 59, PART 5 1 September 1947 No. 335

SOME PHYSICAL ASPECTS OF THE HEAT BALANCE OF THE HUMAN BODY

BY PROFESSOR D. BRUNT, Sc.D., F.R.S.

Presidential Address, delivered 8 *May* 1947

§ 1. THE FACTORS IN THE HEAT ECONOMY OF THE HUMAN BODY

THERE is a continual generation of heat within the body, as a result of digestion and of the oxidation of body tissues associated with physical effort, the rate of generation being called the *metabolic rate.* The heat is conducted to the skin, partly by the normal conduction of the body tissues, partly by the blood stream, and from the skin is dissipated to the environment by the combined action of radiation, convection and the evaporation of sweat. The radiation exchange between the body and its environment will lead to a gain or loss of heat according as the radiating surfaces in the environment are warmer or colder than the skin, and the air will warm or cool the body by convection according as it is warmer or colder than the skin.

It is the temperature of the skin which determines the bodily gain or loss of heat by radiation and convection. While the internal temperature of the body remains nearly constant at about 98° F. in a wide variety of external conditions, the temperature of the skin will vary considerably with variations in these external conditions (temperature, humidity and rate of motion of the air), and will also vary over different parts of the body. Except at air temperatures above 90° F., the skin temperature of a man resting indoors is lowest over the feet, next lowest over the hands, and highest over the trunk and head ; when the air temperature is above 90° F., the skin temperature tends to be uniform or nearly so over the whole of the body (Hardy and DuBois, 1941).

The following notation is employed in accordance with the usual custom:—

M = metabolic rate of generation of heat in the body.

R = rate of radiative loss of heat to the environment, negative when the walls or other radiative surfaces are warmer than the skin.

C = rate of convective loss of heat to the environment, negative when the air is warmer than the skin.

E = rate of loss of heat by evaporation in the lungs and from the skin.

S = storage, or rate of net loss of heat due to lowering of body temperature, counted negative when the body gains heat.

H = rate of gain of heat by absorption of short-wave radiation from sun and sky.

T_a = temperature of the ambient air.
T'_a = wet-bulb temperature of the ambient air.
T_w = temperature of the radiating surfaces near the body.
T_B = temperature of the deep tissues of the body.
T_s = mean temperature of the skin.
T_0 = operative temperature (defined below).
v = speed of air movement in m./sec.
P_s = saturation vapour pressure at temperature T_s.
p_a = actual vapour pressure in the ambient air.

The first six items above measured are in kg. cal. per m^2 of skin per hour.

§2. THE HEAT ECONOMY OF THE BODY INDOORS

In indoor conditions the heat exchange of the body is represented by the equation

$$M+S=R+C+E.$$

Out of doors the absorption of solar radiation, H, and the loss of part of the black-body radiation from solid bodies, are allowed for by adding to M a correction H, and subtracting a quantity which may be 25–30 units with a clear sky in the middle of the day in summer. Of the items in this equation M, S, and E are obtained by direct measurement. M is measured by the oxygen-carbon dioxide exchange in breathing. S is computed from the change in mean temperature of the body, the mean specific heat of the body being estimated to be 0·83. E is computed from the loss of weight of the body in an hour. When these items are known for a variety of conditions, it is possible to compute $R+C$ from the equation above. The experimental data quoted below were obtained at the John B. Pierce Laboratory of Hygiene, Newhaven, Conn., by the Director of the Laboratory (Dr. C.-E. A. Winslow), Dr. L. P. Herrington, and Dr. A. Gagge. The subjects of the experiments were placed in a specially designed booth, in which the temperature, humidity and movement of the air could be controlled, and the temperatures of walls could be controlled by reflecting or heating devices in the walls. The balance used for weighing the subjects was sensitive to 2 gm.

Experiments of this type have usually been limited to nude subjects, since clothing adds a complication to the interpretation of the experiments. The experiments summarized in figure 1 were, however, made with clothed subjects, the clothing consisting of a two-piece suit of cotton underwear, a cotton shirt without tie, socks, leather shoes, and a grey suit with three-quarter-lined coat and fully lined waistcoat. A summary of a considerable number of physiological experiments was given by Brunt (1943).

In figure 1 (Winslow, 1941) the unit is 1 kg. cal. per sq. m. of skin surface per hour, the scale on the right-hand side giving the appropriate total (per hour) for each item for a man of average size, having a skin area 1·8 sq. m., or 19·5 sq. ft. The operative temperature shown as ordinate in figure 1 is a weighted mean of the temperatures of the air and walls, estimated to give a joint representation of the radiation and convection when wall and air temperature are not identical. When these temperatures are identical, the operative temperature is equal to the air temperature.

Figure 1 summarizes observations on clothed subjects resting, with air movement of 17 ft./min. The metabolic rate *M* remains substantially steady in operative temperatures 21 to 30° C. (70 to 86° F.), but increases slowly and steadily as temperature departs above or below this range. The body cools (*S* positive) in operative temperatures below 25·5° C. (78° F.) and is slightly warmed at higher temperatures. The evaporative loss *E* is relatively low at temperatures below 29° C. (84° F.), but increases rapidly as the temperature increases above this limit. Below 84° F. the evaporative loss is entirely due to losses in the respiratory passages and lungs, and to the insensible perspiration which passes through the skin, without wetting the surface. The joint radiative plus convective loss is very great at low temperatures, and is proportional to the difference of mean skin temperature and operative temperature. This item becomes an addition to bodily heat at operative temperatures above 35° C. (95° F.), the air being then warmer than the skin.

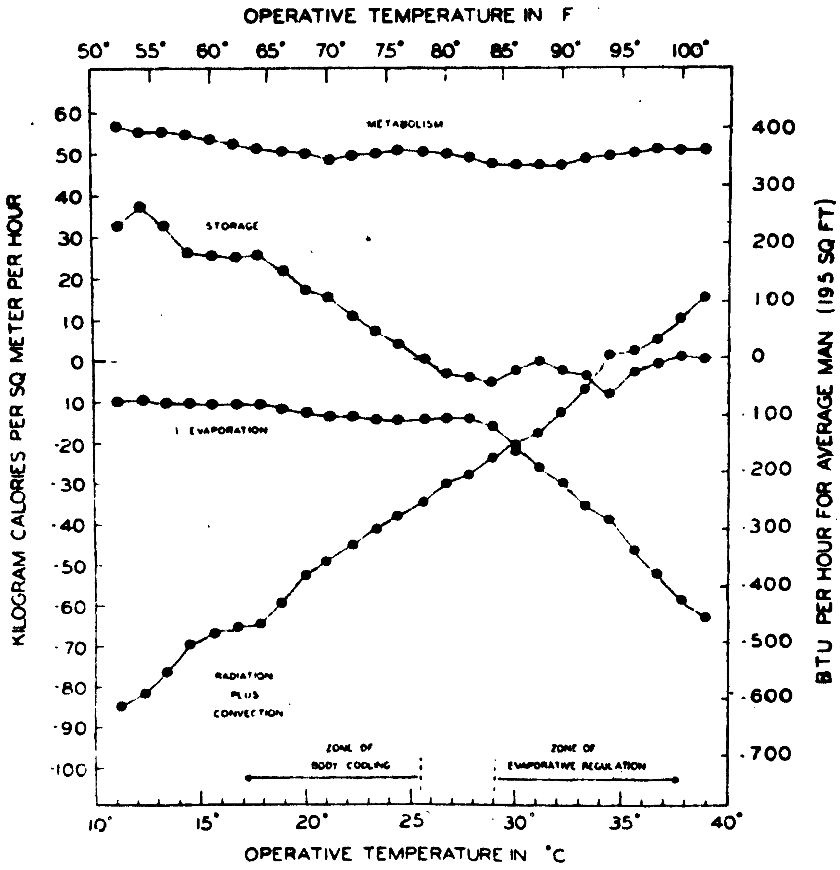


Figure 1. The relation of the different items in the heat exchange of the body to operative temperature. Clothed subjects. (Winslow, 1941.)

Figure 2, which refers entirely to observations on nude subjects (Winslow, 1941), shows, in the lowest curve, the variation of mean skin temperature with changes of operative temperature. Active sweating through the sweat glands sets in at 31° c. for resting nude subjects indoors, instead of at 29° c. as with clothed subjects. The result of this sweating is to maintain the skin temperature at a steady value near 35° c. with operative temperatures above 31° c. With operative temperatures below 31° c. the mean skin temperature falls at a rate of about 1° c. for every 2° c. fall of operative temperature.

The rather irregular curve marked K in the upper part of figure 2 shows the variation of the "conductance" of the body, defined as $K=(M+S)/(T_B-T_s)$. K remains substantially constant at operative temperatures between 31 and 37° c., and increases slightly at temperatures above 37° c. It is also substantially constant at temperatures below 27° c., i.e. in the "zone of body cooling". In the intermediate zone from 27 to 31° c., K increases steadily. This is the "zone of vaso-motor control", in which body temperature is maintained steady by the control of the flow of heat to the skin by the dilation or contraction of the surface blood vessels or, in other words, by the control of the conductance.

It will be noted that when the body is subjected to cold, the "conductance" of heat is decreased by the contraction of the surface blood vessels, while when the body is subjected to heat, above a certain limit the conductance is increased by the complete dilation of the surface blood vessels. It is only in the intermediate zone, known as the "zone of vaso-motor control", and in air temperatures exceeding 37° c., that the degree of dilation or contraction of the peripheral blood vessels is variable. The zone of vaso-motor control is from 25·5 to 29° c. for the clothed body, and from 27 to 31° c. for the nude body, both cases referring to persons resting indoors. An increase in physical activity shifts this zone towards lower temperatures.

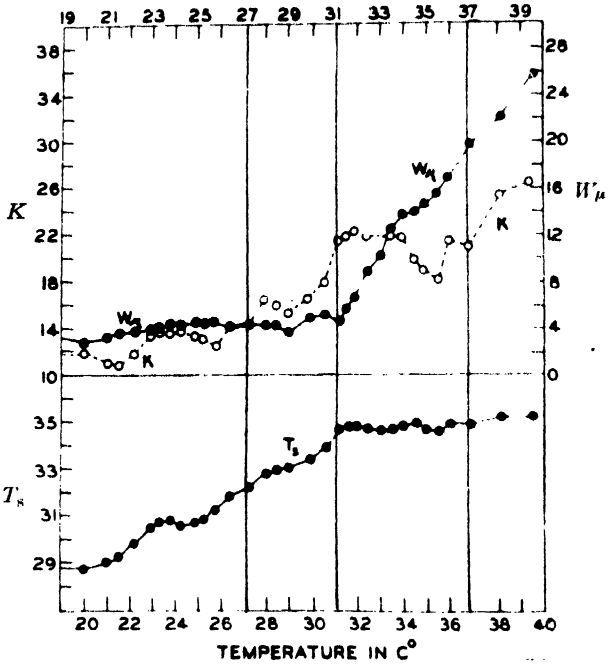


Figure 2. Physiological responses to varying operative temperature. W_μ = wetted area; K = conductance; T_s = mean skin temperature. Nude subjects. (Winslow, 1941.)

Experiments have shown that the conductance is lower in fat than in lean subjects (Winslow, Herrington and Gagge, 1937) and in women than in men (Hardy, Milhorat and DuBois, 1941). Women generally have a more considerable layer of subcutaneous fat than men ; in addition, at low temperatures they have a readier tendency than men to increase their metabolic rate. For these reasons women appear to bear with equanimity a degree of cold which men would find trying were they protected by the same amount of clothing.

The curve marked W in figure 2 gives, in arbitrary units, the area of the part of the skin from which evaporation of sweat is occurring. This area increases rapidly as temperature rises above 31° c. W. is defined as $E/(P_s - p_a)$.

We can now summarize some of the more important reactions of the body to its environment. In a cold environment the surface blood vessels contract, thereby checking the conduction of heat from the deeper tissues to the skin, while the skin temperature falls, thereby decreasing the rate of dissipation of heat from the skin. The metabolic rate is accelerated by the fluid discharged from the adrenal and other glands into the blood stream, when exposure to cold is prolonged. In a warm environment, or during strenuous physical exertion, the peripheral blood vessels dilate, and sweat is freely secreted, the evaporative loss of heat from the skin being the body's main safeguard against rise of temperature.

§3. THE RELATIONSHIP OF RADIATION, CONVECTION AND EVAPORATION TO THE ENVIRONMENTAL CONDITIONS, FOR NUDE SUBJECTS

The observational technique does not separate the quantities R and C. When the temperature of the radiating surfaces, such as walls etc., in the immediate environment, is equal to the air temperature, then for the nude subject

$$R + C = k(T_s - T_a).$$

A number of series of independent experiments showed that k is a linear function of $v^{1/2}$, and that

$$k = 3{\cdot}6 + 16v^{1/2}.$$

It follows that, in general,

$$R = 3{\cdot}6(T_s - T_w),$$

$$C = 16v^{1/2}(T_s - T_a).$$

It can readily be checked that the factor 3·6 for R is in agreement with the assumption that the body radiates as a black body. Various estimates based on experimental determinations agree in yielding for the radiative power of the skin a factor 0·99 of black-body radiation.

The measurements of evaporation (E) represented in figure 1 are those appropriate to a clothed subject reclining, and it is found that the maximum evaporation shown in this diagram is equal to the evaporation from rather less than 30% of the area of skin from which heat is lost by convection. This is not in general agreement with the experience of men in extremely hot conditions, when not reclining. When sawing wood, for example, one finds the whole body becoming wetted with sweat, as it does on exposure to very high temperature.

The curve for evaporation, in figure 1, shows that when the temperature of the ambient air rises above 29° c., the evaporative loss rises, and W in figure 2 shows that the area of skin wetted with sweat increases steadily with rise of temperature. Winslow, Herrington and Gagge (1937) have shown that when the evaporating power of the air is increased by a decrease in humidity, the air temperature being kept constant, there is a decrease in the area of skin wetted by sweat, while the actual loss of heat by evaporation remains unchanged. This is readily seen by a comparison of figures 3 and 4. When the temperature and humidity of the air are kept unchanged, while the air speed is increased, so that the convective loss of heat is increased, the necessary evaporative loss of heat is decreased, and it is also found that the area of skin wetted by sweat is decreased. No instrument has yet been designed which is capable of simulating the human body's control over the area from which evaporation can occur, and it is not surprising that the interpretation of readings of such instruments as have been devised should be uncertain.

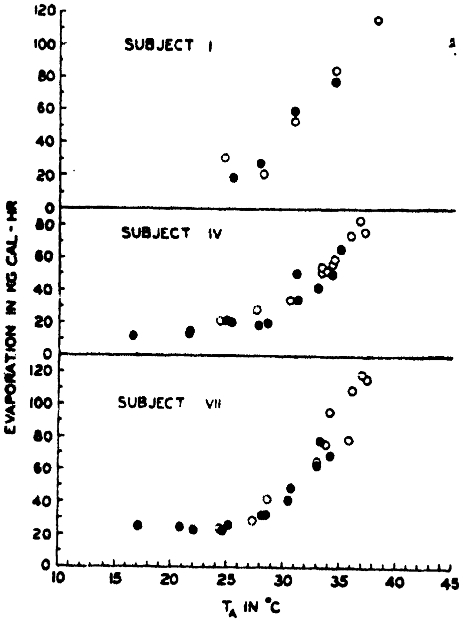


Figure 3. Evaporative heat loss in relation to air temperature for nude subjects. Solid circles, high humidity (40–80%). Open circles, low humidity (14–40%). (Winslow, Herrington and Gagge, 1937.)

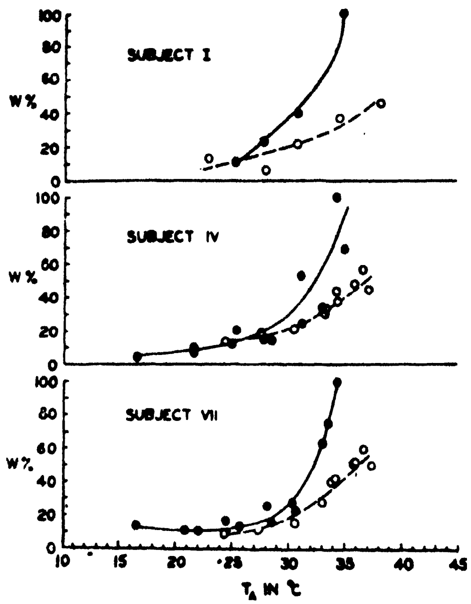


Figure 4. Percentage of maximum possible area of wetted skin surface at varying air temperatures. Nude subjects. Solid circles, high humidity (40–80%). Open circles, low humidity (14–40%). (Winslow, Herrington and Gagge, 1937.)

If we may assume that in the limiting conditions of either great heat or great physical exertion the evaporation of sweat takes place from the same area of skin as the loss of heat by convection, the maximum evaporative loss of heat can be deduced from the convective loss. Convection and evaporation of sweat are then due to air from the environment being brought into contact with the skin, being there brought to saturation at the skin temperature and carried away by turbulent air motion. Thus the heat content and water-vapour content are diffused together, and the result is equivalent to the diffusion of the *total heat content* of air, the latter being the sum of the internal energy of the air and of the latent heat of the water vapour which it carries. If I_s is the total heat content of air saturated at the mean skin temperature, I_a is the total heat content of the ambient air, and I'_a is the total heat content of air saturated at the wet-bulb temperature, then

$$C+E=A(I_s-I_a)=A(I_s-I'_a),$$

for $I_a = I'_a$ is a statement of the ordinary hygrometric equation

$$(c_p + x_a c'_p)(T_a - T'_a) = L'(x'_a - x_a),$$

where x_a and x'_a are the humidity mixing ratios of the ambient air, and of air saturated at the wet-bulb temperature. The humidity mixing ratio is defined as the mass of water vapour mixed with unit mass of dry air. The variation of latent heat with temperature can be neglected as of no importance in the practical problem.

In the equation

$$C + E = A(I_s - I_a),$$

the internal heat of the air, the diffusion of which is the convective loss of heat C, can be readily separated out, giving

$$C = Ac_p(T_s - T_a) = 16v^{1/2}(T_s - T_a).$$

Hence

$$A = 16v^{1/2}/c_p = \tfrac{200}{3} v^{1/2},$$

and

$$C + E = \tfrac{200}{3} v^{1/2}(I_s - I'_a).$$

The function I is only required for saturated air, and a table of its values is given in the Appendix.

§4. THE LIMITING CONDITIONS FOR HEAT-STROKE FOR NUDE SUBJECTS

In a very hot environment, particularly during great physical exertion, the rate of loss of heat by evaporation of sweat will fail to counteract the internal generation of heat plus the heating of the body by convection (when air temperature is above skin temperature). Before this stage is reached, the activity of the sweat glands will have attained its maximum, and the whole body will be wetted with sweat, much of which will drip from the body without evaporating. With continued failure of evaporative loss of heat to balance the internal generation plus convection gain or loss, the body temperature will rise, the pulse-rate will increase, strong palpitation will set in, followed by a condition of stupor; soon afterwards sweating suddenly decreases rapidly and, unless the subject is then immediately removed to a cooler environment, he dies of "heat-stroke".

The limiting conditions of temperature and humidity beyond which heat stroke will come as a result of continued exposure are given, for the *nude* man, by the equation

$$M = 3{\cdot}6(T_s - T_w) + \tfrac{200}{3} v^{1/2}(I_s - I'_a).$$

For given values of M and v the curve which fixes these limits is rapidly drawn. It is necessary to know T_s, the mean skin temperature, but sufficient observations are available to give a reasonably accurate value of T_s for given rates of physical effort. A series of such curves is given in figure 5, in which the ordinate is temperature of the radiating surfaces and the abscissa is the wet-bulb temperature of the air. With given values of M, v, and T_s, the equation is used to deduce I'_a and from this T'_a, corresponding to a series of values of T_w. Figure 5 shows the limiting curves corresponding to the following:—

(1) $M = 47$ (resting):

$T_s = 35{\cdot}6°$ c. $v = 0{\cdot}085$ m./s. Line AA
$v = 1{\cdot}0$ „ „ BB

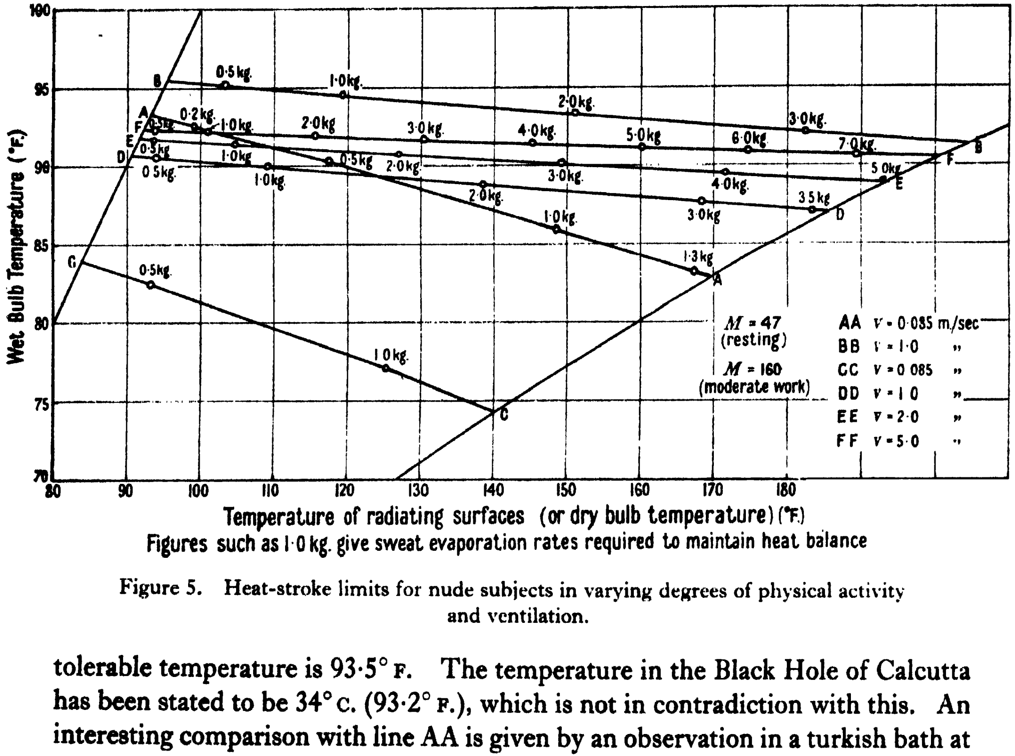
(2) $M=160$ (moderate work):

$T_s=34{\cdot}3^\circ$ C. $v=0{\cdot}085$ m./s. Line CC
$v=1{\cdot}0$ „ „ DD
$v=2{\cdot}0$ „ „ EE
$v=5{\cdot}0$ „ „ FF

Since 1 m./s. = 200 ft./min., the values of v are readily converted into ft./min.

Considering the line AA, we can state that continued exposure to conditions represented by a point above this line, for a man resting in air movement of 0·085 m./sec. (17 ft./min.), will eventually lead to death from heat-stroke. An increase of ventilation raises the tolerable limits, while an increase of physical effort, with its corresponding increase in the internal generation of heat, lowers them.

It would be of interest to obtain some direct check on the form of these curves. For men resting in nearly still saturated air, line AA shows that the maximum



Figures such as 1·0 kg. give sweat evaporation rates required to maintain heat balance

Figure 5. Heat-stroke limits for nude subjects in varying degrees of physical activity and ventilation.

tolerable temperature is 93·5° F. The temperature in the Black Hole of Calcutta has been stated to be 34° C. (93·2° F.), which is not in contradiction with this. An interesting comparison with line AA is given by an observation in a turkish bath at dry-bulb temperature 135° F., wet-bulb temperature 89° F., quoted from J. S. Haldane (1905). The point representing these temperatures is slightly above the limiting line, and it is possible to compute the initial rate of rise of body temperature, which should be 1·0° F. in the 93 minutes the subject of experiment spent in these conditions. The observed rise was 1·7° F., showing that the line AA in figure 5 is a very close representation of the true limits.

For men resting, with air movement 183 ft./min., a close approximation to 200 ft./min., Robinson, Turrell and Gerking (1945) found experimentally the following two limiting conditions:

Dry-bulb temp. 96·8° F. Wet-bulb 96·2° F.
„ „ „ 122° F. „ „ 94·0° F.

McConnell and Houghton (1932) gave 95° F. as the limit in saturated air. The three points thus obtained fit closely the line BB in figure 5.

It thus appears that the lines AA and BB of figure 5 give a reasonable representation of such experimental tests as are available. The factor multiplying $I_s - I_a'$ is rather uncertain for men doing physical work.

In the limiting conditions for any degree of activity and any degree of ventilation, the body is wetted with sweat, and the prevention of rise of body temperature is due to the evaporation. The amounts which must be evaporated from the whole body in the conditions represented by the lines in figure 5 are shown in kilogrammes, at various points along the lines, assuming $T_w = T_a$.

The rate of sweating from the whole body, assuming the skin area to be 1·8 m., is

$$\{M - (3{\cdot}6 + \tfrac{200}{3} v^{1/2})(T_s - T_a)\} \times 1{\cdot}8/L \text{ kg./hr.}$$

When T_w is not equal to T_a, the amount of sweating is increased by an amount

$$\frac{3{\cdot}6 \times 1{\cdot}8}{L}(T_w - T_a) \text{ kg./hr.}$$

In any given conditions it is therefore readily possible to evaluate the rate of sweating required to maintain temperature equilibrium of the body.

When a high rate of sweating is maintained for more than a short time, the loss of water from the body must be compensated, as also must the loss of salt in the sweat. When the rate of sweating exceeds 2 kg./hr., it is not possible to compensate completely for this loss by drinking, as 2 kg./hr. is about the maximum rate at which the body can absorb water. Drinking at a rate exceeding the rate of absorption by the body leads to nausea.

A general comparison of the curves of figure 5 with observation of body temperature of men working at known rates, in known atmospheric conditions, is not possible. It should be emphasized that the value of the conductance, K, in the limiting conditions when the body has become totally wetted with sweat, is not appreciably variable with degree of activity, i.e. with rate of internal generation of heat.

The rate of loss of heat from unit area of skin is given by

$$M + S = K(T_B - T_s).$$

Suppose a man is exposed to conditions in which, with a low value of M, this equation is satisfied with $S = 0$, and that he starts working at a rate which increases his total output of heat to a new value M'. Then

$$M' > K(T_B - T_s).$$

It will not be possible to attain thermal balance of the body until $K(T_B - T_s)$ is increased, either by an increase of T_B or a decrease of T_s, or an increase of K. Clearly the immediate reaction is an increase of K, if this is possible; but if M' is much greater than M, then $T_B - T$ must increase before thermal balance is attained, since K is not increased by even 100% in the range of environmental temperatures shown in figure 2..

In general, an increase of $T_B - T_s$ appears to be attained partly by a rise of T_B and partly by a fall of T_s. But the initial rise of T_B, which is the commonly observed physiological variable, does not indicate that the conditions make an eventual thermal equilibrium impossible.

§5. THE GENERAL EFFECTS OF TEMPERATURE AND HUMIDITY

Figure 6 gives in AA and BB a conversion of curves AA and BB of figure 5 into a diagram in which the coordinates are dry-bulb temperature and relative humidity, assuming the radiating surfaces to be at air temperatures. Both curves give the limiting conditions for a nude man at rest indoors, AA for air movement 17 ft./min., and BB for air movement 200 ft./min. CC gives the limiting conditions in which body equilibrium can be maintained by a clothed man resting in sunshine with only about one-third of his skin wetted by sweat, and DD the limiting conditions for a clothed man walking 3 miles per hour in bright sunshine, with the same area of skin wetted by sweat. Figures 5 and 6 are expressed in degrees Fahrenheit for convenience of comparison with observations made in industry and in meteorology.

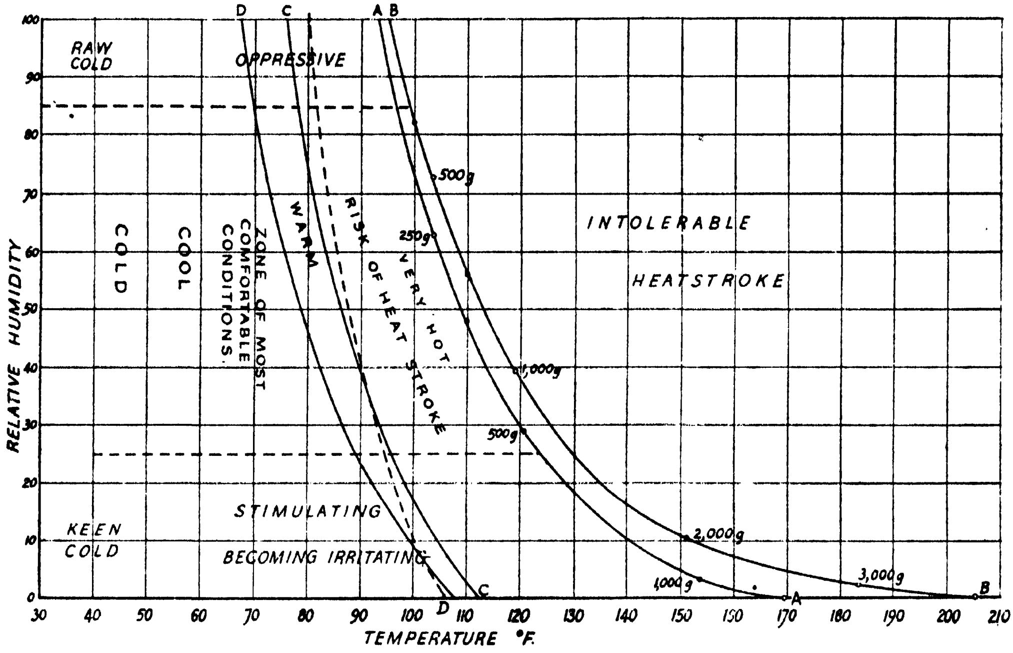


Figure 6. Tentative classification of climates.

AA. Heat-stroke limits for nude man resting in still air.

BB. Heat-stroke limits for nude man resting in air moving 200ft./min.

CC. Limiting conditions for clothed man resting in sunshine with about one-third of skin wetted with sweat.

DD. Limiting conditions for clothed man walking 3 m.p.h. with about one-third of skin wetted with sweat.

The broken line represents equivalent temperature 80° F. The figures 500 g., etc., indicate rate of evaporation of sweat in grammes per hour for men of average size in order to maintain heat balance of the body.

The character of the sensation experienced in conditions represented by different areas in the diagram is shown by the words printed in the diagram. The effect of high humidity is to give, in cold air, the feeling of cold usually called "raw", and in warm air a feeling of oppression. With low humidity, cold air gives that impression of cold which is usually called "keen", while warm, very dry air is at first stimulating, but ends by producing irritation and quarrelsomeness, as exemplified by the Föhn wind, a warm dry wind which is frequently observed blowing down mountain slopes. Dwellers in hot desert regions manifest an unpleasant permanence of irritability.

The optimum conditions for human beings fall within the middle ranges of both temperature and humidity. With relative humidity of 50–60%, the optimum temperature is between 60 and 76° F., in reasonably good agreement with the recommendations of the heating and ventilating engineers. For air-conditioned rooms, the indoor temperature recommended in the U.S.A. is 76° F. in summer, and 72° F. in winter, and in Great Britain 66° F. in summer and 62–64° F. in winter.

Brunt (1943) has given a specification of the ideal outdoor climate on the assumption that this should permit a man, lightly clothed, (*a*) to walk at 3 miles per hour in sunshine without sweating, and (*b*) to rest in bright sunshine or to stand in the shade or indoors doing light work, with light air movement (17 ft./min.), without sweating or body cooling. The optimum for these prescribed conditions is 67° F., for a lightly clothed man, when the relative humidity is near 50%. For the *nude* man a slightly higher temperature, perhaps 70° F., is required.

§ 6. A COMPARISON WITH NATURAL CLIMATES

Figure 6 would exclude very cold and very hot climates, and also very damp and very dry climates, as undesirable for permanent residence. The most desirable climates are those having moderate temperature and humidity, and those are found in the middle latitudes of the globe. In the tropics the monotony of the weather from day to day is the most marked feature and is experienced even at tropical places so far above sea-level as to have temperatures which are not extreme.

Figures 5 and 6 are computed on the assumption that the heat balance of the body is determined by the rate of internal generation of heat, radiation, convection and evaporation of sweat, and figure 6 involves the further assumption that the radiating surfaces in the environment are at air temperature. The effect of very hot radiating surfaces can be great, and may be evaluated as shown earlier. In hot, sunny climates the walls and roofs of houses, and of any solid objects exposed to sunshine, all attain very high temperatures, and they not only radiate heat to the body at a higher rate than is allowed for in figures 5 and 6, but they are extremely uncomfortable to the touch. In these conditions an air temperature of 105° F. is about as high as a white man can tolerate.

There are two types of hot natural climates—the warm, damp, cloudy and oppressive type, and the hot, dry and sunny type. If air conditioning were possible, we should treat the first type by *drying* the indoor air, and the second by *cooling* it. Failing this, all that can be done in the warm damp climate is to increase ventilation, or to sit in the greatest draught, and in the hot, dry climate to close all doors and windows early in the morning, so as to exclude the entry of hot air from outside, while increasing ventilation by electric fan.

In warm, damp climates houses should be of an open type, permitting free air movement, and the clothing should be similar in character, being light and porous. In hot, sunny climates houses should have double walls and roof, with wide air spaces between, and should be capable of being closed tightly against the intrusion of hot air from outside during the day. Clothing should be such as to reflect as much as possible of the solar radiation falling upon it, and should be loose and not very permeable. It should be added that white cotton of two thicknesses reflects 71% of incident radiation, while a khaki shirt reflects only 43%. A white skin reflects about 45%, a dark brunette skin 35%, and a negro's 16%.

Climatic conditions represented by points to the right of the line CC in figure 6 are considered unsuitable for permanent settlement by the white man. Allowing for the variation of temperature from day to day and from year to year in any given month, we arrive at the conclusion that, if the relatively frequent occurrence for 3–4 hours per day of conditions hotter than correspond to the line CC is to be avoided, the mean temperature of the hottest month should not exceed 75° F. in a dry climate, or 73° F. in a damp climate, if the wind is light at the hottest hours of the day. It is desirable that an analysis of, say, African climates should be made on this basis.

§7. ACCLIMATIZATION AND SETTLEMENT IN THE TROPICS

Within certain limits it is possible for a man to accustom himself to living and working in higher temperatures than can at first be tolerated by a newcomer. This acclimatization consists partly in "training" the sweat glands to function readily and economically, partly in a loss of weight, with a corresponding increase in conductance and of the ratio of skin area to total weight of the body, and partly in learning to avoid unnecessary physical exertion. After acclimatization the regulation of body temperature is more efficient, and the rise of body temperature following physical exertion is markedly less than it is initially. How precisely acclimatization is brought about is obscure, and it is found that during the later stages of acclimatization the usually observed physiological variables, such as rectal and skin temperatures, pulse-rate and rate of breathing, show a stable reaction from day to day. An interesting illustration of this will be found in Horvath and Shelley's (1946) description of their experiments in acclimatization of men to work in rather extreme conditions of heat and humidity. The work in question consisted in marching at about 3 miles per hour, carrying a 20-lb. pack. On the first day they marched for ½ hour only; on the next 13 days they marched for 1 hour. At the end of the march the rectal temperature, mean skin temperature pulse rate and sweat loss were measured. The rise of each of the first three factors all showed marked decrease from day to day during the first six days, and during the subsequent days maintained a stable level. The sweating rate increased during the whole 14 days. The men showed very marked improvement in gait, ease of working and general appearance during days 7–14, although no change was noted in the measured physiological variables. It is clear that in the later stages of acclimatization there occur physiological changes of an undetermined nature, which mark an improvement in bodily well-being but do not lend themselves to direct measurement.

Both laboratory experiments and experience in the tropics suggest that strenuous work or sport should form part of the daily life of any white man who desires to be healthy in the tropics. In Queensland, for example, it is found that the woman who does hard physical labour, such as scrubbing offices, remains healthy, while the sedentary worker, like the typist, deteriorates in health, and the woman who does nothing becomes sickly.

Professor D. H. K. Lee of the University of Queensland has written that a normal consequence of the acclimatization of the body to hot damp climates is a decline of the will to work in the absence of stimulus. If not corrected, this leads to reduced output of work or to carelessness, in either case reducing morale.

A summary of a very great volume of literature dealing with settlement in the tropics has been given by Grenfell Price (1939). A perusal of Price's book shows that there are wide divergences of opinion as to the possibilities of white men settling in the tropics. Many medical men appear to hold the view that, with the conquest of such diseases as malaria and hookworm,we shall have made it possible for men to reside in any part of the tropics. Experience has shown that while the white man can settle in the marginal tropics, i.e. the trade-wind coasts and islands and the tropical plateaux, the settlement of the inner tropical regions, where both temperature and humidity are high, cannot be made without loss of both physical and mental efficiency.

ACKNOWLEDGMENT

Thanks are due to the Royal Meteorological Society for the loan of blocks of figures 1, 2, 3 and 4.

REFERENCES

Brunt, D., 1943. *Quart. J. R. Met. Soc.*, **69**, 77.

Haldane, J. S., 1905. *J. Hygiene*, **5**, 494, Expt. XII.

Hardy, J. D. and DuBois, E. F., 1941. " Temperature, its Measurement and Control in Science and Industry" (Symposium, *American Inst. of Phys.* (New York: Reinhold Publ. Co.), p. 537.

Hardy, J. D., Milhorat, A. T. and DuBois, E. F., 1941. *Temperature, its Measurement, etc.*, p. 529.

Horvath, S. M. and Shelley, W. B., 1946. *Amer. J. Physiol.*, **146**, 336.

McConnell, W. J. and Houghten, F. C., 1923. *Jour. A.S.H.V.E.*, **29**, 131.

Price, A. Grenfell, 1939. " White Settlers in the Tropics." *Amer. Geog. Soc.*, Special Publ., No. 23.

Robinson, S., Turrell, E. S. and Gerking, S. D., 1945. *Amer. J. Physiol.*, **143**, 21.

Winslow, C.-E. A., 1941. *Temperature, its Measurement, etc.*, p. 509.

Winslow, C.-E. A., Herrington, L. P. and Gagge, A. P., 1937. *Amer. J. Physiol.*, **120**, 288.

APPENDIX I

Table of values of the total heat content of saturated air

T (° c.)	*I*	*T* (° c.)	*I*	*T* (° c.)	*I*	*T* (° c.)	*I*
0	2·3						
1	2·7	11	7·5	21	14·2	31	24·1
2	3·1	12	8·1	22	15·0	32	25·3
3	3·5	13	8·7	23	15·9	33	26·6
4	4·0	14	9·3	24	16·8	34	28·0
5	4·5	15	10·0	25	17·7	35	29·4
6	5·0	16	10·6	26	18·7	36	30·9
7	5·5	17	11·2	27	19·7	37	32·4
8	5·9	18	11·9	28	20·8	38	34·0
9	6·4	19	12·6	29	21·9	39	35·8
10	6·9	20	13·4	30	23·0	40	37·7

I is given in kilogramme-calories per kilogramme of dry air plus the water vapour it contains.

APPENDIX II

Value of M for man working

For a man working at the rate of 1000 W ft./lbs. per minute the rate of generation of heat which has to be dissipated from the skin, expressed in kg. cal. per sq. m. of skin per hr. is given to a sufficient degree of approximation by

$$M=64+35W.$$

This assumes that the efficiency of the human body regarded as a heat machine is about 24%.

SOME INVESTIGATIONS IN THE FIELD OF HEAT TRANSFER

By MAX JAKOB,
Illinois Institute of Technology, Armour Research Foundation,
and Purdue University

Thirtieth Guthrie Lecture, delivered 3 *October* 1946

§ 1. INTRODUCTION

May I begin with thanks to the Council of the Physical Society for the high honour extended to me, first in 1937 and now again, by the invitation to give the lecture in commemoration of the founder of this Society, Professor Frederick Guthrie. In 1937, being just about to leave Europe, I was not able to accept this invitation. I can scarcely express how happy I am to have been given another opportunity to deliver the Guthrie Lecture. I understand that Mrs. Guthrie, until her death about two years ago, did not miss any of the lectures in honour of her late husband, and I am pleased that Professor Guthrie's daughter and other relatives of his are with us tonight in continuation of this tradition of their family.

Thinking of my former stays in London, I recall with particular gratitude the friendship extended to me by two prominent members of your Society; these were Sir Richard T. Glazebrook, whom I met first in 1929 when he was President of the First International Steam Tables Conference, and then when we had to deal with the international heat unit, and Professor H. L. Callendar, whose merits in flow calorimetry, resistance thermometry, and steam research I emphasized in a talk at the banquet of the Steam Tables Conference. I made friends with other members of your Society when giving a series of lectures on steam research at the University of London in 1931; for instance, Dr. Ezer Griffiths, your former Honorary Secretary, to mention only one of them. I also had the opportunity of attending the presentation of the Duddell Medal for

1930 to Sir Ambrose Fleming and was deeply impressed by that meeting at which the late Sir Arthur Eddington was in the chair.

Re-reading the list of Guthrie Lecturers, I feel that it was less because of my own small contributions to science that I was inserted in this brilliant list, but rather in appreciation of an engineer's work in some fields of science less cultivated by the physicists.

I often gave thought to the question why the science of heat transfer is so far in arrears compared to its sister sciences, optics, electricity, and fluid dynamics. Concerning heat radiation, the main difficulties may be found in the badly defined surfaces, complex geometrical configuration and temperature distribution, and in unexpected anomalies of the absorption by gas mixtures. The progress in experimental research on problems of heat conduction was so much slower than in corresponding work on electrical conduction because of the following facts :

1. The ubiquity of large temperature differences in nature, causing parasitic heat currents, whereas environmental electric potential differences will seldom influence electric conduction problems.
2. The difficulty of heat insulation, whereas electrical insulators for moderate voltage are usually almost perfect. This is particularly bad because of the existence of temperature differences, just mentioned.
3. The slowness of temperature propagation compared with the enormous velocity of electricity. This, of course, is due to the inclination of matter to store up heat energy instead of carrying all of it forward.
4. The complex configurations which have to be dealt with.

Considering finally heat convection, the situation becomes even worse. The Biot-Fourier equation of thermal conduction has to be combined with Stokes's equations of viscous flow; variable properties of the flowing medium, badly defined boundary conditions, and the phenomenon of unstable transition between laminar and turbulent flow have to be taken into consideration.

These facts made work in heat transfer difficult, inexact, cumbersome, and therefore unattractive. In fact, scientists were deterred from dealing with such unwieldy stuff, promising a scarce and belated crop only, not worth the labouring.

So it happened that other branches of science flourished and the branch of heat transfer developed only very slowly. Hence our knowledge of the processes in the interior of a molecule and even in the nucleus of an atom is better than our knowledge of the heat flow in a cup of tea, let alone in an industrial furnace.

I would have liked to draw a more detailed, but general picture of the field of heat transfer with its entangled roots of radiation, conduction, and convection and of special tools used in its cultivation, such as similarity and analogy methods. However, as your President, Professor D. Brunt, has so graciously mentioned, I had to prepare for this lecture at short notice and, therefore, I can show you only a few fruits from this field, just as I have them at hand from work done in recent years in cooperation with some younger colleagues. They may give you an idea of the variety of problems which pass the laboratory and desk of a specialist in heat transfer.

§ 2. HEAT RADIATION

(a) *A derivation of the basic law of gas radiation*

In work on the selective absorption of radiation by gases I started with a new derivation of the law of emission of radiation by a gas body of arbitrary shape. Koenigsberger's (1903) derivation for a rectangular parallelopiped is general, but rather clumsy. Simpler derivations have been presented for a sphere, for instance, by Nusselt (1923); however, as Koenigsberger mentioned, they are not general because arbitrary bodies cannot be built up from elementary spheres. The following proof is simple and general. It follows a usual pattern, but seems not to have been published previously.

Let* V be a volume of finite size and arbitrary shape (figure 1), containing an absorbing gas, and let S be the perfectly black inner surface of a spherical shell with centre C, inside V. The radius R of the sphere may be so large that the linear dimensions of the space V can be considered as differentials compared with R. The substance between S and V is to be non-absorbing, but have the same index of refraction as the gas in V. The whole system is to be kept in thermal equilibrium at the absolute temperature T.

All radiation from surface element dS that enters the volume V is contained in the solid angle ω_V. A differential part of it, in the solid angle $d\omega_V$, may cross the volume V in a cross-sectional area dA which, owing to our assumption concerning the size of V, can be considered as constant along the distance L. For the same reason the beam from dS to V is assumed to be perpendicular to dS.

Then the time rate of the radiation of wave-length λ which enters V in the solid angle $d\omega_V$ will be $d^2q = N_{b\lambda} \,.\, d\omega_V \,.\, dS$, where $N_{b\lambda}$ is the monochromatic areal radiant intensity for a black body.

The amount absorbed in V is

$$d^2q_\alpha = N_{b\lambda} \,.\, d\omega_V \,.\, dS(1 - e^{-m_\lambda L}), \qquad \ldots\ldots(1)$$

where m_λ is the logarithmic decrement of radiation for wave-length λ.

Since L is of differential magnitude compared to R, it can be taken so small that $m_\lambda L \ll 1$ and therefore $m_\lambda{}^2 L^2 \ll m_\lambda L$, so that $e^{-m_\lambda L} \rightarrow 1 - m_\lambda L$.

Further, by definition, $d\omega_V = dA/R^2$. Hence, from equation (1), $d^2q_\alpha = N_{b\lambda} \times (dA/R^2) dS \,.\, m_\lambda L = N_{b\lambda}(dS/R^2) m_\lambda \,.\, dV$.

By substitution of $dS/R^2 = d\omega_S$ and double integration (over S and V), $q = 4\pi m_\lambda N_{\gamma b} V$.

Since gases can be considered as non-reflecting, Kirchhoff's law requires the time rates of absorbed and emitted energy, q_α and q_ϵ, to be equal. Herewith the law of heat emission is obtained in the form used by Koenigsberger, $q_\epsilon = 4 m_\lambda W_{b\lambda} V$ where $W_\lambda = \pi N_{b\lambda}$ is the monochromatic radiant flux density for a black body.

(b) *A photographic and photometric model method for the determination of surface and gas radiation*

Though the laws of emission and absorption of radiation by surfaces and absorbing gases are well known (see § 1 (a)) straight analytical determination

---

* In general the symbols recommended by the American Standards Association (1943) will be used in this lecture.

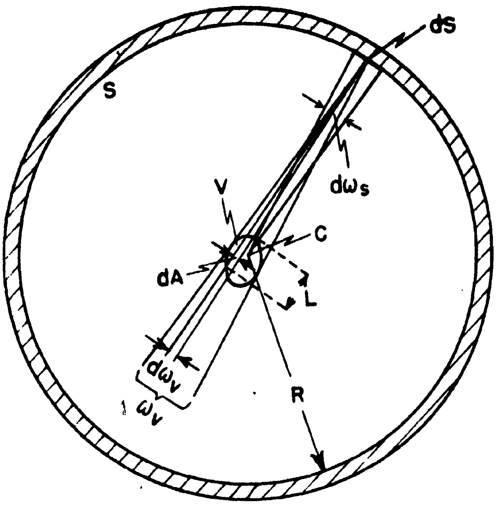


Figure 1. Radiation of a gas body.

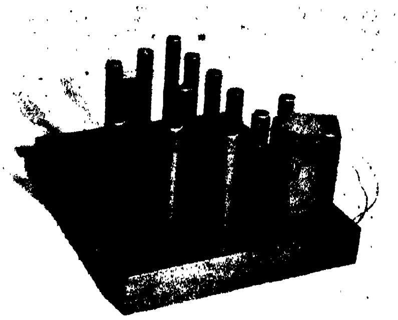
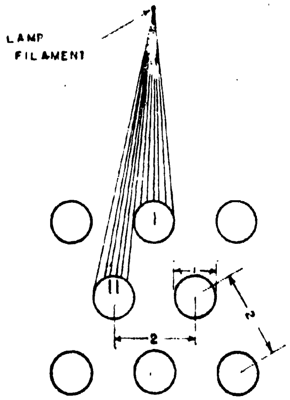


Figure 2. Model of a bank of tubes with lamp house.



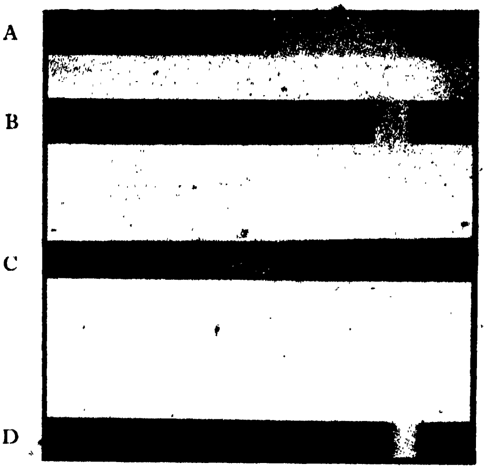


Figure 4. Spectrograms (A: lamp; B: lamp and filter ; C : lamp and liquid ; D : lamp, filter and liquid).

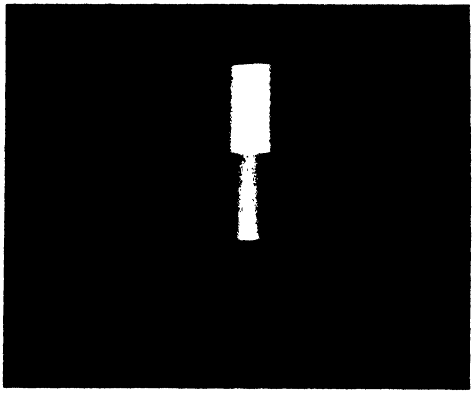
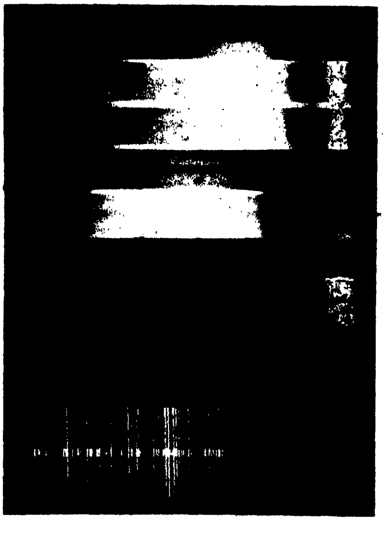


Figure 5. Spectrograms for three grades of the solution (top : weak ; middle : medium ; bottom : strong).



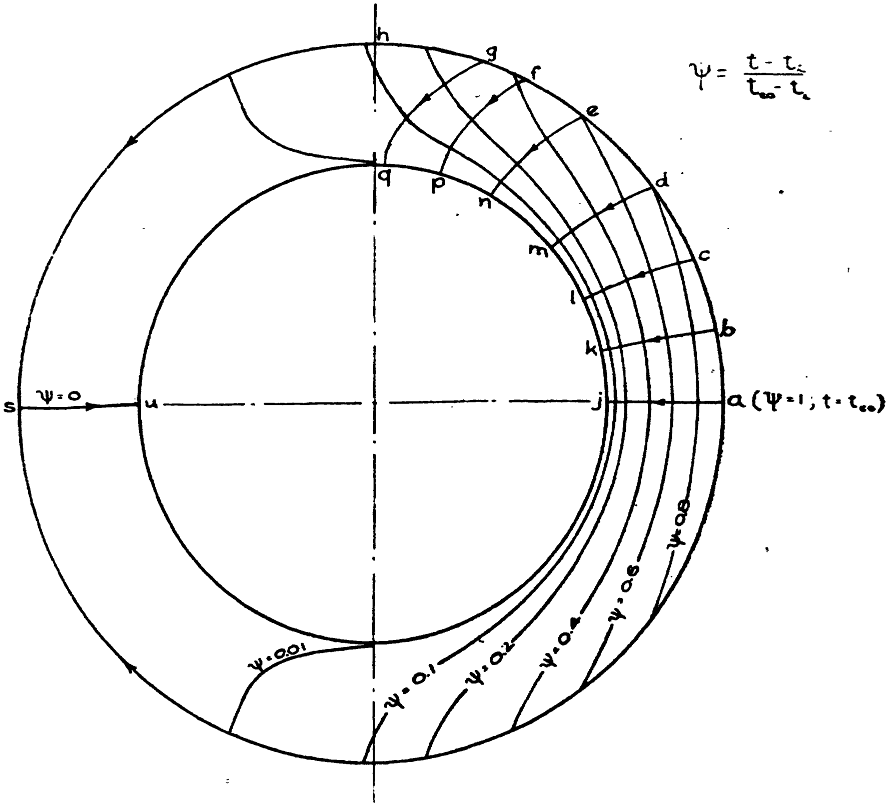


Figure 8. Isothermals and flow lines in boiler-tube wall.

of the heat exchange by radiation in a device is possible only if the emitting and absorbing surfaces are very simple in shape and configuration. In most practical cases approximative calculations and graphical methods have been used in order to evade the mathematical difficulties. In this section, an experimental model method will be described which has been developed in cooperation with Professor Hawkins of Purdue University (Jakob and Hawkins, 1942). A small model, similar to the device under investigation, was constructed. Light was used instead of the long-wave heat radiation, and its absorption by photographic films which covered the surfaces of the model was measured by the methods of photometry. Liquids which absorb light in a narrow wave-length range were used instead of absorbing gases.

Figures 2 and 3 represent the model used. It consisted of a lamp-house and a number of pipes, 1 inch outer diameter, 10 inches long, mounted on a base plate. The light source was a single-filament galvanometer lamp which was located in the lamp-house or without that in an adjustable mounting. By horizontal slits in the housing, radiation could be restricted to a horizontal plane. When used without housing, the whole filament, about 4¾ inches long, acted as a source of radiation, parallel to the tube axes. With films wrapped around the pipes the device was ready for exposure.

Calibration consisted in exposing sections of the films for various lengths of time, developing the films and passing the resulting negatives through a recording photometer. The intensity of light transmitted through the clear film base was used as reference intensity in determination of the photographic density after different times of exposure.

The blackening of the films by exposure in the model was also measured with the recording photometer. It is apparent from figure 3 that the blackening was not uniform around the tubes. Tube I, for instance, causes a shading effect on tube II.

The complex case of absorption of radiation by gases was imitated by immersing the entire apparatus in water containing a dye. It was essential to use a combination of light source and filter which transmits light having a wave-length range in which the liquid transmits. This was not the only restriction, since it was also necessary to select a film which is sensitive to the wave-length range of light which is transmitted through the absorbing medium. The method would be useless if the film selected was " blind " to the radiation coming from the combination of light source, filter and absorbing medium.

After a great deal of experimentation, satisfactory results were eventually obtained using various concentrations of saffron dye in distilled water, Wratten light filters, and Wratten and Wainwright "hyperpanchromatic" cut films.

Figure 4 shows a series of spectrograms made with a Hilger spectroscope. The continuous spectrum of the lamp is shown in spectrogram A. Spectrogram B is taken with lamp and filter, C with lamp and liquid, showing that the liquid transmits light in almost the same range as does the filter. Spectrogram D is for light passing through filter and absorbing liquid.

In order to change the absorbing power of the medium, it was sufficient to change the concentration of the dye. Figure 5 shows the same as the spectrogram D for three grades of the absorbing solution.

Figure 6 contains eleven spectrograms which are explained by table 1.

Table 1. Explanation of figure 6

Spectrogram No. (from top to bottom)	Source of light and absorbing bodies	Time of exposure (seconds)
1	Lamp	10
2	Lamp	30
3	Lamp	90
4	Lamp and liquid	90
5	Lamp and liquid	300
6	Lamp and filter	90
7	Lamp and filter	300
8	Lamp, filter and liquid	180
9	Lamp, filter and liquid	600
10	Iron arc	3
11	Iron arc	3

Imitation of complex radiating surfaces may be obtained by suitable distribution of several filaments on the surface or, more simply, by bringing one filament to different places on the surface, exposing at each position, and superimposing the results. A variety of other possibilities, as well as of difficulties, has been indicated in the quoted publication.

§ 3. HEAT CONDUCTION

(a) *Temperature distribution in the walls of boiler tubes*

In a boiler the outer surfaces of the steam tubes will be at a much lower temperature than other surfaces in the furnace. It is not easy to determine the temperature of these tube surfaces. Using optical pyrometers the reflexion of radiation which they receive must be carefully considered. Otherwise enormous errors may occur, as has been shown in previous papers (Reid and Corey, 1944; Jakob, 1944).

The present example (Jakob, 1943 a) deals with the determination of the outside temperature of boiler tubes by an indirect theoretical method, based on a close estimate of the distribution of the radiation outside and on calorimetric measurements of the increase of enthalpy of the fluid inside the tubes. These measurements were performed in a high-pressure boiler with forced circulation by Davidson, Hardie, Humphreys, Markson, Mumford, and Ravese (1943). The theoretical calculation further yielded information about the temperature distribution in the tube walls and about the thermal resistance between the tubes and the boiling water.

Figure 7 is a cross-section through three adjacent boiler tubes. The heat is assumed to arrive at a steady rate from the right (furnace side). The left side (wall side) is supposed to be virtually insulated by the boiler wall. It can be assumed that the main part of the heat energy arrives as radiation either from all directions of the furnace space with uniform intensity or in the direction OU. In the first case the time rate of heat radiation to a surface element at any point P of tube A is proportional to a solid angle ω which itself is proportional to the ordinary angle ψ; in the second case it is proportional to $\cos\phi$. A mathematical analysis showed that the two assumptions yield almost the same distribution of absorbed energy. Neglecting, for the time being, the heat exchange by radiation between tubes A and B, it is as though heat sources existed at the surface which decrease in strength as $\cos\phi$ from $\phi=0$ to $\phi=\pi/2$.

The thermal conduction through the tube wall under the steady-state condition is governed by the partial differential equation

$$\frac{\partial^2 t}{\partial r^2}+\frac{1}{r}\frac{\partial t}{\partial r}+\frac{1}{r^2}\frac{\partial^2 t}{\partial \phi^2}=0$$

where t is the temperature at the radial distance r from the centre line and the angular distance ϕ from the direction UO.

For a complete circular ring (dashed area of tube A in figure 7), a solution of the differential equation was obtained by means of Fourier series. In putting up the boundary conditions the following physical assumptions were made :

$$t=t_i=\text{constant} \quad \text{for } r=r_e,$$

$$\left(\frac{\partial t}{\partial r}\right)_{r=r_e}=\frac{q_0''}{k}\cos\phi \quad \text{for } 0\lesseqgtr\phi\lesseqgtr\pi/2,$$

$$\left(\frac{\partial t}{\partial r}\right)_{r=r_e}=0 \quad \text{for } 0\lesseqgtr\pi/2\lesseqgtr\pi,$$

where

q_0'' = the heat energy * absorbed in unit time and area at $\phi=0$,

k = the thermal conductivity of the wall material, and

t_i = a temperature slightly above that of the saturated steam t_{sat} inside the tube.

Both k and t_i are considered to be constant.

Denoting by t_{e0} the temperature at the point 0 ($r=r_e$, $\phi=0$) a temperature ratio may be defined by $\Psi=\frac{t-t_i}{t_{e0}-t_i}$ having values between 0 (for $t=t_i$) and 1 (for $t=t_e$).

For the case of $r_e/r_i=1{\cdot}5$ the temperature distribution shown in figure 8 was obtained by numerical calculation. This figure contains isothermal lines for $\Psi=1$, 0·8, 0·6, 0·4, 0·2, 0·1, 0·01, and 0. Further, heat-flow lines are drawn from the points a, b, c, d, e, f, g, and h on the outside surface to j, k, l, m, n, p, q, and u on the inside surface, respectively. Each of the first four channels, beginning with the symmetry line aj, carries 1/5 of the total heat flow, the channel between en and fp carries 1/10, each of the two last channels carries only 1/20 of the heat flow. It is further seen that about 96% of the incoming heat is received in the range $\phi=0$ to $0{\cdot}4\pi$ at the outside and is given up inside in the first quadrant ($\phi=0$ to $0{\cdot}5\pi$) and only 4% in the rear quadrant. The temperature difference t_e-t_i at $\phi=\pi/2$ is about 12% of $t_{e0}-t_i$ (at $\phi=0$).

After some corrections for radiation exchange between tubes A and B and some other radiation due to the actual arrangement in the tests, a satisfactory agreement between theory and experiment was obtained regarding the values of t_{e0}. Other surface temperatures had not been measured because of the

---

* I use prime, double prime, and triple prime signs to designate unit length, area, and volume, respectively.

difficulties with thermocouples to be placed on tubes in a high-pressure steam boiler.

An additional qualitative result of this analysis was that inside the tube at $\phi = 0$ local coefficients of heat transfer of the order of 20000 or 30000 B.hr.$^{-1}$ ft.$^{-2}$F.$^{-1}$ occur, and in the range from $\phi = 0$ to $\pi/2$ mean coefficients of the order of 10000 or 20000 B.hr.$^{-1}$ft.$^{-2}$F.$^{-1}$. The high local heat transfer comes from the strong formation of steam bubbles at the front side which induces vehement radial and rotational movements of the mixture in the fluid cross-section, and from the wiping effect of fluid forced through the tubes. The rear half of the inner tube surface, on the other hand, is not engaged at all in the heat transfer; it acts solely as heat protection of the furnace wall. As an average, the investigation showed that the thermal resistance of the fluid film amounts only to about 10% of the total resistance, that is, the resistance against thermal conduction in the tube wall is the controlling factor.

(b) *Temperature distribution in electrical coils of simple form due to a linear increase of Joulean heat with temperature.*

Whereas § 3 (a) dealt with a case of conduction of heat which was carried to the conducting body from outside by radiation and was carried away by convection, it will now be supposed that heat is developed in the conducting body.

An electrical coil is a rather inhomogeneous body because it is built up from conducting and insulating materials. However, considering equal volumes of such size that several layers of these materials are included, it can be assumed that in each volume the same Joulean heat is developed if the electric resistance of equal lengths of the conductor is the same all through the coil.

The heat developed in such a body is conducted to the surface as it would be in a homogeneous medium whose thermal conductivity is equivalent to that of the mixture of materials in the coil. In a steady state of heat development and heat flow, the temperature t decreases from a maximum value t_0 somewhere inside the coil to lowest values t_s at the surface. It will be assumed that t_s is uniform all over the coil surface. This holds approximately for a coil in an oil bath or in a fast gas stream and in numerous other cases.

Simple relations between maximum and mean temperature of a coil have been derived for the case of uniform generation of heat all over the volume. In particular, it was found that for coils, having the shape of an infinitely wide plane plate, an infinitely long cylinder, or a sphere, the temperature distribution may be expressed by

$$\theta = \theta_0(1 - \xi^2), \qquad \ldots\ldots(2)$$

where $\theta = t - t_s$ is the temperature excess over surface temperature for a point at the perpendicular or radial distance, $x = s$ or $r = s$ from the median plane, centre line, or centre of the plate, cylinder, or sphere, respectively; $\theta_0 = t_0 - t_s$ is the same for $x = 0$ or $r = 0$; s is the half thickness or the radius of the coil; and ξ is defined as the ratio x/s or r/s.

Owing to the increase of the electric resistance with temperature, the heat developed in a coil is not uniform, but increases from the surface to the point of maximum temperature. Having previously dealt with the case of uniform development in a coil whose cross-section is a rectangle of finite side lengths (Jakob, 1919), I later studied the influence of non-uniform development of heat in coils of simple shape in ranges where the electric resistance can be assumed to increase linearly with temperature (Jakob, 1943 *b*). The general procedure and some surprising results of this study will be dealt with.

In addition to the above mentioned symbols, the following will also be used:

k = equivalent (apparent) thermal conductivity of the combination of electrically conducting and insulating material as used in the coil;

q''' = rate of heat energy developed in unit volume;

$$\frac{q'''}{k}=m+n\theta=m\left(1+\frac{n}{m}\theta\right)=m(1+\epsilon\theta), \qquad \ldots\ldots(3)$$

where $m=q_s'''/k_s$, $n=\epsilon q_s'''/k_s$, $\epsilon=n/m$ = temperature coefficient, $\sigma=s\sqrt{n}$. Subscripts s, 0, and m refer to the places of surface, maximum, and mean temperature, respectively.

For the infinitely wide plate, the differential equation of the temperature distribution is

$$\frac{d^2\theta}{dx^2}=-\frac{q'''}{k}.$$

For $q'''>0$ and $n>0$, its general solution is

$$0=-\frac{m}{n}+M\cos(x\sqrt{n})+N\sin(x\sqrt{n}),$$

where M and N are the constants of integration.

Boundary conditions are

$$d\theta/dx=0 \quad \text{when} \quad x=0, \text{ and } \theta=0 \quad \text{when} \quad x=s.$$

Using them if follows that

$$\theta=\frac{1}{\epsilon}\left[\frac{\cos(\xi\sigma)}{\cos\sigma}-1\right], \qquad \ldots\ldots(4)$$

$$\theta_0=\frac{1}{\epsilon}\left[\frac{1}{\cos\sigma}-1\right], \qquad \ldots\ldots(5)$$

and

$$\frac{\theta}{\theta_0}=\frac{\cos(\xi\sigma)-\cos\sigma}{1-\cos\sigma}. \qquad \ldots\ldots(6)$$

Thus, the temperature distribution across the plate is not parabolic, as in equation (2) which holds for $\sigma=0$, but is more complicated. When σ approaches $\pi/2$, then θ approaches infinity for every value of x. Hence, if equation (3) were valid up to this limit, every point of the plate would be at infinitely high temperature. θ/θ_0, however, would keep the finite value

$$\frac{\theta}{\theta_0}=\cos\left(\xi\cdot\frac{\pi}{2}\right). \qquad \ldots\ldots(7)$$

The temperature distribution for $\sigma=0$ and $\sigma=\pi/2$ is represented in figure 9; it changes surprisingly little in the whole range from $\sigma=0$ to $\sigma=\pi/2$ (infinitely high temperature).

The mean temperature is obtained from the equation of definition

$$\theta_m = \frac{1}{s}\int_0^s \theta \, . \, dx.$$

Substituting from equation (4) and integrating,

$$\theta_m = \frac{1}{\epsilon}\left(\frac{\tan\sigma}{\sigma} - 1\right). \qquad \ldots\ldots(8)$$

From equations (5) and (8) one obtains the ratio

$$\phi = \frac{\theta_m}{\theta_0} = \frac{\sin\sigma - \sigma \, . \cos\sigma}{\sigma(1-\cos\sigma)}. \qquad \ldots\ldots(9)$$

This equation can be used to calculate the maximum temperature excess θ_0 from θ_m which is easily determined by measuring the increase of the coil resistance when carrying current.

For $\sigma = 0$, that is, uniform heat development, $\phi = 2/3$. For the other limit ($\sigma = \pi/2$) equation (9) yields $\phi = 2/\pi$, that is, only $4\frac{1}{2}$% less, independent of the thickness of the plate and the value of the temperature coefficient ϵ.

For cylindrical and spherical coils similar equations were derived to those for the plane plate.

The only difference in the solution for the cylinder is that the equations contain the Bessel function of zero order and first kind, J_0, wherever the cosine-function occurs in the equations above, and infinitely high temperature would take place where the Bessel function becomes zero, i.e. at the first zero point of J_0 which occurs for $\sigma = j_{0,1} = 2{\cdot}4048\ldots$

For the sphere the cosine-functions in equations (4), (5), and (6) have to be replaced by $[\sin(\xi\sigma)]/(\xi\sigma)$, including $(\sin\sigma)/\sigma$ and the temperature approaches infinity when $\sigma \rightarrow \pi$.

The temperature distributions (figures 10 and 11) cover a band which is only slightly wider than the one between $\sigma = 0$ and $\sigma = \pi/2$ in figure 9.

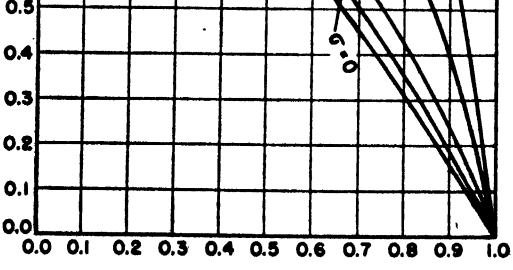


Figure 9. Temperature distribution in infinitely wide plates.

The upper limit ($\sigma = 0$) of this band is in all cases the parabolic distribution (equation (2)), the lower limit is given by equation (7) for the plate,

$$\theta/\theta_0 = J_0(\xi \, . j_{0,1}) \text{ for the cylinder,} \qquad \ldots\ldots(10)$$

and

$$\theta/\theta_0 = [\sin(\xi\pi)]/(\xi\pi) \text{ for the sphere.} \qquad \ldots\ldots(11)$$

The practically important ratio $\phi = \theta/\theta_0$ takes maximum values 0·667, 0·5, and 0·4 (for uniform heat development) and minimum values of 0·637, 0·432, and 0·304 (at infinite temperature) for plane plate, cylinder and sphere, respectively.

Obviously, engineers neglecting in their calculations the influence of non-uniformity of heat generation in a coil, were just lucky; they could scarcely have foreseen that a relatively small increase in current would theoretically lead to infinitely high temperature and yet to such small influence upon the ratio ϕ.

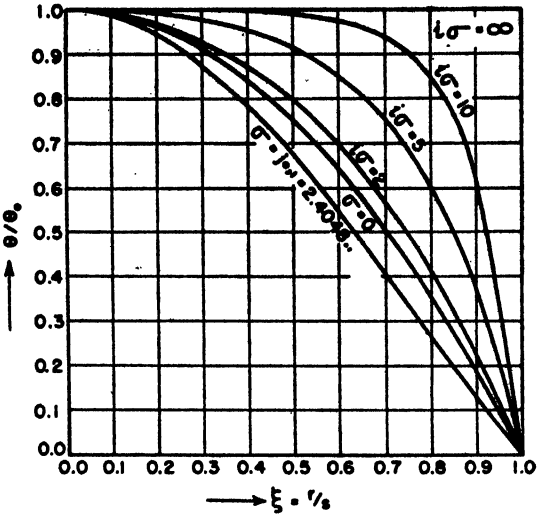


Figure 10. Temperature distribution in infinitely long cylinders.

(c) *Temperature distribution in bodies of simple form developing or absorbing heat at a linear function of temperature*

The relations for the temperature distribution in electrical coils given in § 3 (b) are also valid for the case of chemical exothermic reactions in a range where a linear increase of the heat development with temperature can be assumed.

It is easily understood that they will hold also for endothermic reactions ($q''' < 0$; $m < 0$) in the range where equation (3) is valid and $n > 0$, that is, for heat absorption increasing with decreasing temperature. In particular, θ/θ_0 remains entirely unchanged; however, θ, θ_0, and θ_m assume negative values, as they must be in endothermic reactions.

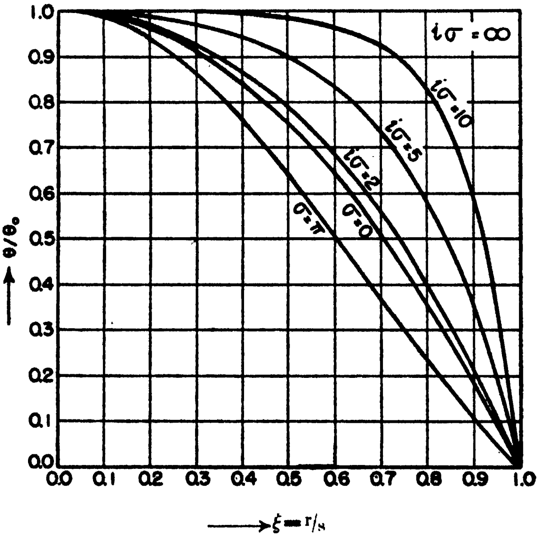


Figure 11. Temperature distribution in spheres.

Also the limit of an infinite temperature value, shown in figures 9, 10, and 11, would be theoretically the same. However, just as a coil would burn through, or equation (3) would cease to be valid long before an infinitely high temperature were obtained, any endothermic reaction would stop before the absolute zero point of temperature were attained.

Similar relations to those given in § 3 (b) have recently been derived (Jakob, 1947) for electric heat sources with negative temperature coefficient of the Joulean heat, i.e. heat development decreasing linearly with increasing temperature, as may occur in the carbon of electric-arc lamps or graphite electrodes, or in electrolytes; these equations then are also valid for exothermic reactions with negative temperature coefficient, and for endothermic reactions with $n<0$, that is, when the absorption of heat increases with temperature. In the latter case, again, θ assumes negative values.

The procedure of derivation is similar to that shown in §3(b). In general, the results, including those of §3(b), may be represented by two equations:

$$\frac{\theta}{\theta_0}=\frac{f(\xi\sigma)-f(\sigma)}{1-f(\sigma)}, \qquad \ldots\ldots(12)$$

and

$$\frac{\theta}{\theta_0}=\frac{g(i\xi\sigma)-g(i\sigma)}{1-g(i\sigma)}, \qquad \ldots\ldots(13)$$

where $i=\sqrt{-1}$ and f and g are function symbols. Considering, for instance, $f(\sigma)$ and $g(i\sigma)$, the first one means $\cos\sigma$, $J_0(\sigma)$, and $(\sin\sigma)/\sigma$, and the second one $\cosh(i\sigma)$, $I_0(i\sigma)$, and $[\sinh(i\sigma)]/(i\sigma)$ for the plate, cylinder and sphere, respectively. The function f belongs to exothermic and endothermic processes with positive n, the function g to the same with negative n; generation of Joulean heat is included in the exothermic processes. The argument $i\sigma$ is always real and positive. In figures 9, 10 and 11 the families of curves above the parabola which is common to plane, cylinder and sphere belong to equation (13). They approach an upper limit, $\theta/\theta_0=1$, when $\sigma\rightarrow\infty$.

§4. HEAT CONVECTION

(a) *Studies on free convection*

The Nusselt and Grashof numbers for the free convection on a vertical surface of height H and temperature t_s to a fluid of temperature t_m are usually defined by $(N_{\mathrm{Nu}})_{\mathrm{H}}=\frac{hH}{k}$ and $(N_{\mathrm{Gr}})_{\mathrm{H}}=\frac{\beta g}{\nu^2}H^3(t_s-t_m)$, where h is the film coefficient of heat transfer and k, β and ν are the thermal conductivity, coefficient of thermal expansion, and kinematic viscosity of the fluid, respectively.

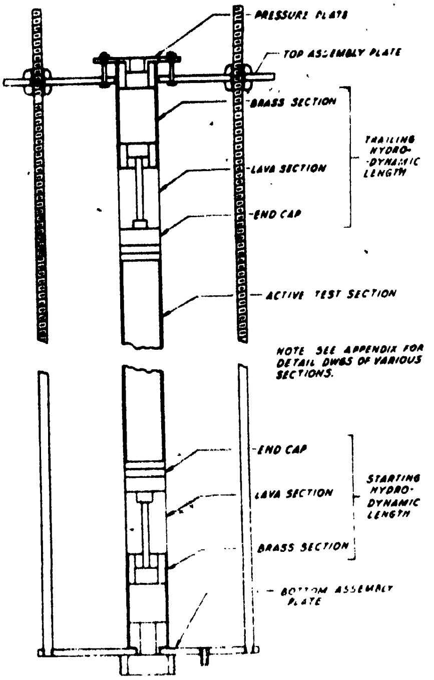
Employing the principle of similarity, Nusselt (1915) showed that $(N_{\mathrm{Nu}})_{\mathrm{H}}$ is proportional to $H^{3/4}$, a result which previously had been obtained analytically by Lorenz (1881). This, however, is only a fair approximation to the actually much more complicated relation between h and H, as has later on been demonstrated by different workers. Moreover, Griffiths and Davis (1922) have shown by experiments that above a certain height (about 2 feet for air) h becomes independent of H and proportional to t_s-t_m. They correctly concluded that turbulence occurs above that height.

Considering the general form

$$N_{\mathrm{Nu}}=C(N_{\mathrm{Gr}}\,.\,N_{\mathrm{Pr}})^n, \qquad \ldots\ldots(14)$$

where N_{Pr} is the Prandtl number, a correlation of King (1932) proved that $n=\frac{1}{4}$ and $\frac{1}{3}$ occur for the laminar and turbulent range, respectively. This holds also for not too vehemently boiling water according to experiments of Jakob and Fritz (1931) and Jakob and Linke (1933, 1935).

The latter workers, in particular, showed that $n=\frac{1}{3}$ may be used for the convection on the upper side of a horizontal plate on which water is boiling, and they concluded that in the range of their experiments ($q''=7$ to 5200 B.hr.$^{-1}$ ft.$^{-2}$), the coefficient of convection on a horizontal plate is independent of the size



NOTE SEE APPENDIX FOR DETAIL DWGS OF VARIOUS SECTIONS.

Figure 12. Vertical-cylinder arrangement for free-convection experiments.

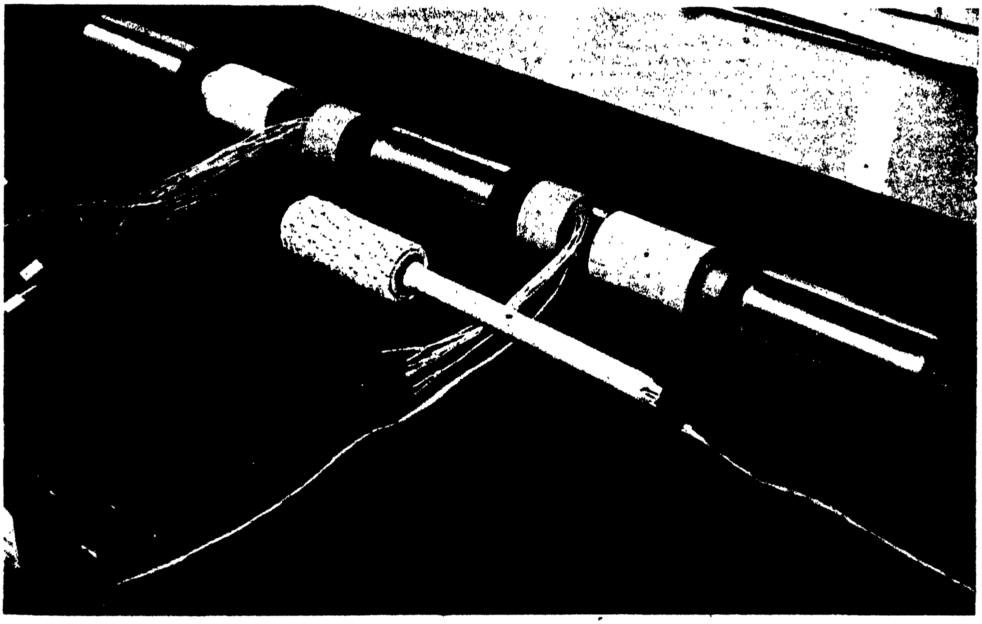


Figure 13. Parts of apparatus for free-convection experiments.

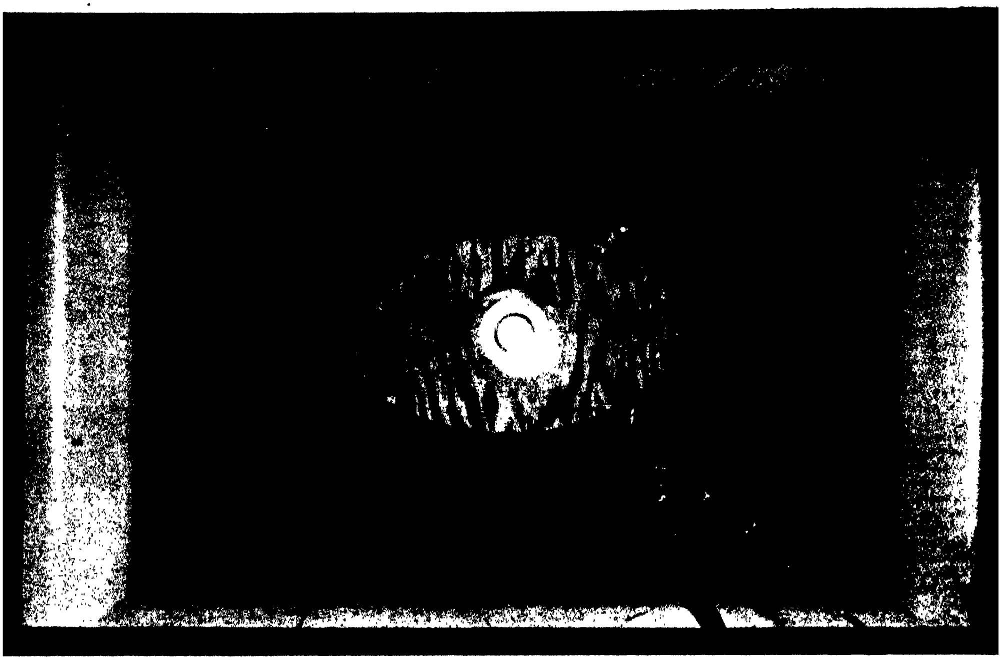


Figure 26. Arrangement of Calorimeter No. 2 in wind tunnel (front side).

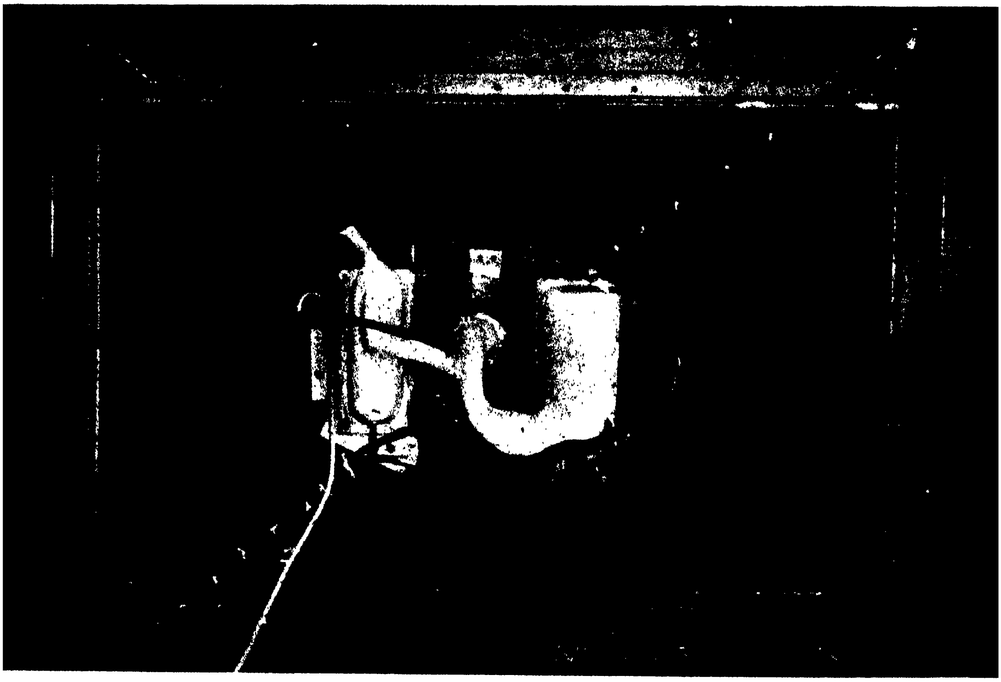


Figure 27. Arrangement of Calorimeter No. 2 in wind tunnel (rear side).

of the plate except for a possible effect of the edges. This was to be expected because there is no reason why the heat transfer on one place of a large horizontal plate should be different from that at any other place.

In other words, if equation (14) is valid, then $n=\frac{1}{3}$ must necessarily occur in all cases where no reason exists why h should depend on a characteristic length.

The heat transfer by free convection from vertical cylinders to non-boiling liquids (water and ethylene glycol) has recently been investigated by Touloukian, Hawkins, and Jakob (1947). Long cylinders, 2¾ inches in diameter, were composed from heating elements, starting and trailing pieces as is shown in figures 12 and 13. Three sizes of heating elements, 6, 12, and 36 inches long, but always the same starting piece (12 inches long) and trailing piece (13 inches long) were employed in the experiments. The cylinder assembly was placed in a cylindric shell of 12 inches inside diameter which contained the liquid.

The test sections consisted of brass tubes containing electrical heaters. Surface temperatures were measured on different places by means of thermocouples in slots which were cut into the surface and were afterwards closed by strips of lead. A number of thermocouples were distributed in the liquid bulk. Fifteen thermocouples inside each end-cap (see figure 12) served to determine the heat losses through the caps. These were made of Transite board, an insulating cement-asbestos compound.

The lava and brass sections (figure 12) served for mounting the test section and as hydrodynamic starting and trailing pieces.

In order to prevent the liquid from penetrating the joints of the cylindric test section and to make the porous Transite and lava sections impermeable to liquids, several thin coats of a special resin were applied to the surface with a spray gun and baked on with a battery of infra-red lamps.

The range of experiments is shown in table 2.

Table 2. Range of experiments about free convection in liquids on vertical surface

Item	Symbol	Minimum	Maximum	Units*
Height of heating section	H	0·5	3·0	ft.
Surface temperature	t_s	90·5	239·5	°F.
Temperature difference	$t_s - t_m$	3·5	83·5	F.
Coefficient of heat tranfer by convection	h	17·6	151·0	B.hr.$^{-1}$ ft.$^{-2}$ F.$^{-1}$
Nusselt number	N_{Nu}	89·0	903·0	—
Grashof number	N_{Gr}	2·2(10^6)	326(10^9)	—
Prandtl number	N_{Pr}	2·4	117·8	—
Product of Grashof and Prandtl numbers	$N_{\mathrm{Gr}} \cdot N_{\mathrm{Pr}}$	280(10^6)	904(10^9)	—

The tests in the laminar region led to an exponent $n=\frac{1}{4}$ in equation (14) as would be expected. In the turbulent region, however, equation (14) had to be replaced by

$$N_{\mathrm{Nu}} = C(N_{\mathrm{Gr}} N_{\mathrm{Pr}}^{m})^{n}, \qquad \ldots\ldots(15)$$

---

* For several years the author has used the symbols C and F as units of temperature *differences* in the Centigrade and Fahrenheit scales, and the symbols °C. and °F. for temperatures in these two scales. This distinction seems to be useful and is recommended for general adoption.

with $m = 1{\cdot}29$ and $n = 1/3$ in order to correlate the results by a unique line. This indicates that the term of acceleration in Stokes' equations cannot be entirely neglected as is done when the same exponent is given to N_{Gr} and N_{Pr}.

The correlations are represented in figures 14 and 15.

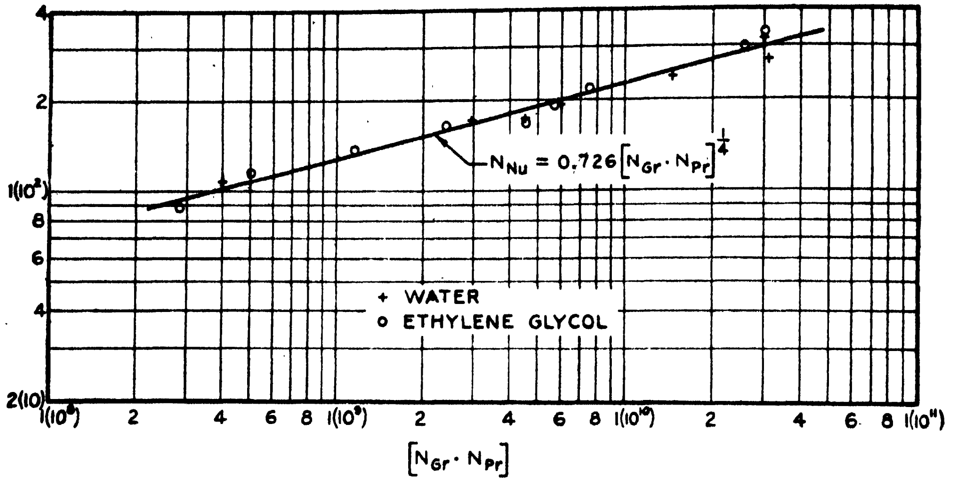


Figure 14. Heat transfer by free convection of liquids on vertical cylinders (laminar range).

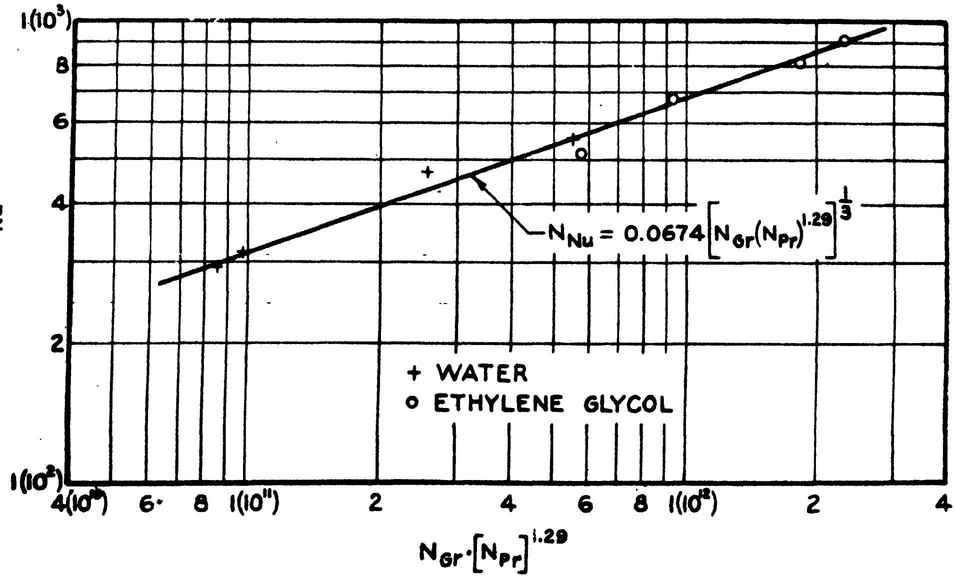


Figure 15. Heat transfer by free convection of liquids on vertical cylinders (turbulent range).

The characteristic exponents 1/4 and 1/3 were also applied in a new correlation of experiments on free convection through enclosed plane gas layers (Jakob, 1946). Mull and Reiher (1930), in a remarkable experimental investigation, had used two parallel plates, each 40 inches long and 24 inches wide, which were separated by air layers. These could be divided in different ways so that it was possible to study enclosed air layers having areas of 30 to 960 square inches. Air layers of seven different thicknesses from ½ to 7¾ inches were employed. One of the two plates was an electrical heating plate and the arrangement was made in a manner to ensure that exactly measured amounts of heat flowed across the air layers. The system of plates could be rotated in a horizontal bearing so that the air layers could be brought into horizontal, vertical, or oblique position. The authors introduced the concept of an equivalent conductivity, k_c, which includes the effect of conduction and convection through the air layer (after deduction of the effect of radiation) and studied the ratio k_c/k in which k is the true thermal conductivity of the air. This ratio, as is well known, is identical with a Nusselt number UL/k where

$$U = \frac{1}{1/h_1 + L/k + 1/h_2} = \text{overall coefficient of heat transfer.}^*$$

h_1, h_2 = film coefficients of heat transfer on the two surfaces, and L = the thickness of the air layer.

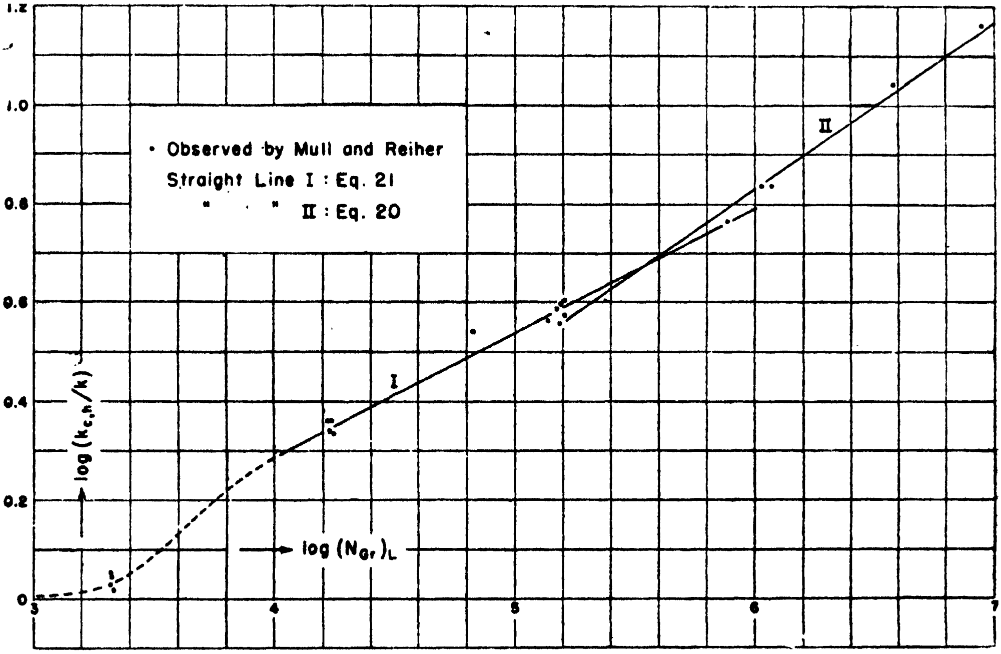


Figure 16. Heat transfer by free convection in horizontal air layers.

For horizontal layers (subscript h) with heat flow upward, Mull and Reiher represented their results by plotting $k_{c,h}/k$ versus $\log(N_{Gr})_L$, a Grashof number with the layer thickness L as characteristic length. In the experimental range of $(N_{Gr})_L = 2\,100$ to $8\,890\,000$ a smooth curve of increasing steepness was obtained which could not be represented by a simple formula. Representation in bilogarithmic coordinates, however (figure 16), reveals that in close relationship to other cases of free convection, two ranges must be distinguished and can be covered by two equations.

---

* After deduction of radiation.

In a lower range $(N_{Gr})_L \approx 10000$ to $(N_{Gr})_L \approx 400000$:

$$k_{c,h}/k = 0{\cdot}195(N_{Gr})_L^{1/4}, \qquad \ldots\ldots(16)$$

In the turbulent range $(N_{Gr})_L > 400000$:

$$k_{c,h}/k = 0{\cdot}068(N_{Gr})_L^{1/3}. \qquad \ldots\ldots(17)$$

For $(N_{Gr})_L \to 0$, finally:

$$k_{c,h}/k \to 1, \qquad \ldots\ldots(18)$$

as indicated by a dotted line.

The meaning of equation (17), obviously, is that above a certain thickness L the coefficient of heat transfer does not change any more, but remains the same as for a single horizontal plate, facing upward. This was checked numerically and found to be in excellent agreement with an equation derived from the observations of Griffiths and Davis.

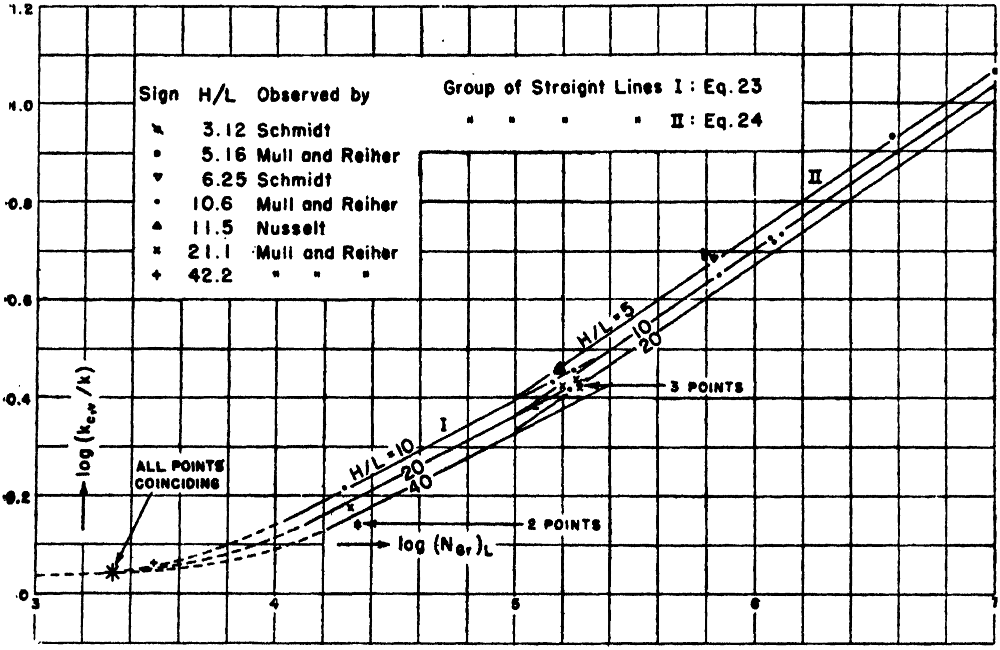


Figure 17. Heat transfer by free convection in vertical air layers.

For vertical layers (subscript v) of thickness L and height H, Mull and Reiher plotted $(k_{c,v}/k)(H/L)$, as measured, versus log $(N_{Gr})_L$ and built up 27 curves with H/L as parameter, making ample use of interpolation and extrapolation since a total of only 21 points from their own and 3 from other experiments were available. Again, bilogarithmic plotting (figure 17) of the original points yields a much simpler and more reliable picture and two simple equations, namely for $(N_{Gr})_L = 20000$ to 200000:

$$k_{c,v}/k = 0{\cdot}18(N_{Gr})_L^{1/4}(H/L)^{-1/9}, \qquad \ldots\ldots(19)$$

and for $(N_{Gr})_L = 200000$ to 10000000:

$$k_{c,v}/k = 0{\cdot}065(N_{Gr})_L^{1/3}(H/L)^{-1/9}. \qquad \ldots\ldots(20)$$

Again, the exponents 1/4 and 1/3 of $(N_{Gr})_L$ are significant for laminar and turbulent flow. The exponent $-1/9$ of H/L, however, was found empirically. At $H/L=0$, equations (19) and (20) would yield an infinitely large heat transfer, whereas the behaviour of single vertical plates should be approached; however, it seems that equation (20) is valid down to the ratio $H/L \approx 3$.

For oblique layers, linear interpolation between the results of equations (17) and (20) gives reasonable results.

Finally it must be mentioned that Mull and Reiher's tests, as well as the above equations, do not take account of any convection or conduction effect of the border strips which close the gas layers.*

(b) *Heat transfer to a fluid in laminar flow through an annular space*

The experimental part of this investigation (Jakob and Rees, 1941) was performed with an arrangement originally constructed and used for the determination of the true temperature and the heat exchange in a catalytic reaction (Jakob, 1938 and 1939). A sketch of the annulus with the thermocouples used is shown in figure 18, in which the distances in the length direction are represented in the right proportions; those in radial direction are magnified and not to scale. The annulus, 1500. mm. long, was formed by two vertical co-axial nickel tubes T_1 of 4·05 mm. o.d., and T_2 of 8·1 mm. i.d. A thin-walled steel tube containing a fine thermocouple could be shifted up and down inside T_1. On the outside of tube T_2 eleven fine iron-constantan thermocouples were fixed. Another thirteen couples were placed on the outer surface of a third nickel tube T_3 of 21 mm. o.d. A fourth tube T_4 of brass, 30·2 mm. o.d., was fitted with three heating coils, each covering a third of the tube length and provided with separate current control. Air, hydrogen, or ethylene were passed through the annulus and heated electrically so that the temperatures of the tubes T_1 and T_2 increased as linearly as possible over the length of Section II. Then a convection-heat coefficient h_{21} was found which was defined by the equation

$$VC_p \,.\, \Delta t = h_{21} \,.\, 2\pi r_2 \,.\, L \,.\, \delta t,$$

where $\dot{V}$ = the time rate of volume flow, C_p = the specific heat of unit volume of the gas at constant pressure, Δt = the temperature increase of the gas over a length L, r_2 = the inner radius of tube T_2, and δt = the average temperature difference between tubes T_2 and T_1.

Up to about 2% of the heating energy crossed the annulus in the form of radiation and was then supplied to the gas from the inner wall. Average gas temperatures were from 31 to 105° c. with temperature slopes from 0·15 to 0·80 c./cm. and differences δt from 0·4 to 12·9 c. The Reynolds number N_{Re} was varied from 50 to 1000; it was defined by

$$N_{Re} = \frac{\dot{V}D_e}{\nu A}. \quad D_e = 2(r_2 - r_1) = \text{equivalent diameter of annulus.}$$

r_1 = the outer radius of tube T_1. ν = the kinematic viscosity of the fluid.
$A = \pi(r_2^2 - r_1^2)$ = the cross-sectional area of the annulus.

---

* Only after delivery of this lecture was a thorough experimental and theoretical paper of Elenbaas (1942) brought to the author's attention, in which the heat transfer from *both* surfaces to air layers *open* at the perimeter is treated. This case is somewhat related to the one dealt with above. It was not possible to compare the results and include an analysis of the comparison in this lecture.

The experimental results were compared with the theory in which laminar flow and invariable physical properties of the fluid were assumed. The derivation was based on the known equation of the velocity distribution across an annular space and on the heat-flow balance for a volume differential of the annulus. This led to the differential equation

$$\frac{\partial}{\partial r}\left(r\frac{\partial t}{\partial r}\right)=Nr\left(r_1^2-r^2+B\ln\frac{r}{r_1}\right)\frac{\partial t}{\partial x}-r\frac{\partial^2 t}{\partial x^2}$$

where $N=\dfrac{2\dot{V}C_p}{\pi Mk}$, $M=(r_2^2-r_1^2)(r_2^2+r_1^2-B)$ and $B=\dfrac{r_2^2-r_1^2}{\ln(r_2/r_1)}$.

In the case of uniform heating or cooling of the fluid from either or both boundary surfaces of the annulus, it can be supposed that the heat energy is absorbed or given up by the fluid in such a manner that at any sufficiently large axial distance x from the entrance and at any radial distance r from the centre line the temperature increase per unit length is constant, so that

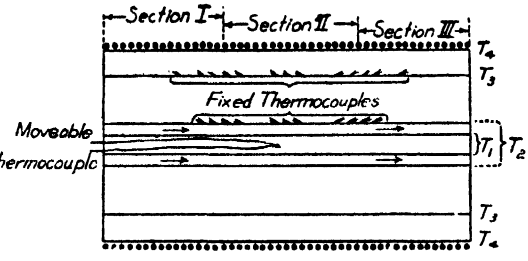


Figure 18. Sketch of annulus arrangement.

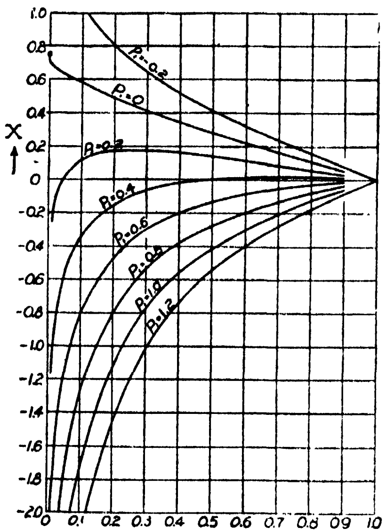


Figure 19. Shape factor X according to equation (21).

$$\frac{\partial t}{\partial x}=C=\text{constant, and } \frac{\partial^2 t}{\partial x^2}=0.$$

Let the time rates of heat entering the annulus from the inner and outer surface (subscripts 1 and 2) be q_1 and q_2, which can be different in numerical value and sign; further,

$$q=q_1+q_2 \quad\text{and}\quad P_1=q_1/q,\quad\text{whilst}\quad P_2=1-P_1=q_2/q.$$

Then for temperatures t_1 and t_2 and for any value x:

$$\begin{aligned}\frac{t_2-t}{CL}&=\frac{\delta t}{\Delta t}=\frac{\dot{V}C_p}{2\pi kL}\frac{1}{(1-y^2)(1+y^2-z)}\{(1-y^2)[\tfrac{3}{4}(1+y^2)-z]\\&\quad+y^2(y^2-z)\ln y+[1-z-y^2(y^2-z)]P_1\ln y\}\\&=\frac{\dot{V}C_n}{2\pi kL}\mathrm{X}=\frac{1}{2\pi}N_{\mathrm{Gz}}\,.\,\mathrm{X},\qquad\qquad \text{......(21)}\end{aligned}$$

where $y=r_1/r_2$ and $z=\frac{y^2-1}{\ln y}$, N_{Gz}=the Graetz number, and X is a function, both defined by equation (21).

By this equation it is possible to calculate the ratio of the temperature difference across the annulus, δt, to the temperature increase along the annulus, Δt, from the Graetz number and the term X which is also dimensionless.

Values of this term are represented in figure 19 and, according to equation (21), it is a linear function of fraction P_1 which is used as parameter in the chart. Since, by definition, $P_1+P_2=1$, the inequality $P_1>1$ indicates that some of the heat which is supplied to the fluid from the inner tube is given up to the outer tube and from this to the environment as occurs in heat exchangers consisting of three co-axial tubes. In the inverse case, $P_1<0$, heat is given up to the inside. It is easily understood that for $r_1/r_2 \to 0$ the function X will approach $\pm\infty$ for every finite value of P_1, because the transfer of a finite amount of heat by an infinitely thin wire requires an infinitely great temperature drop. Only for $P_1=0$ a finite value $X=0{\cdot}75$ occurs (see figure 19). For $r_1/r_2 \to 1$, on the other hand, $X \to 0$.

Making allowance for some experimental difficulties explained in the paper, the experimental results obtained with one annulus and three gases of very different thermal conductivities were in reasonable agreement with the theory.

(c) *Forced heat convection in laminar and turbulent flow parallel to a surface*

Heat transfer between a surface and air flowing parallel to it is a process of great practical importance which, for instance, occurs with all kinds of fins or on the skin of an airplane in flight. Previous knowledge of that process was based on a few sets of experiments which were performed with plane surfaces and led to a considerably higher heat transfer than a theory due to Latzko (1921). In particular, the influence of non-heated starting sections seemed to require a new investigation. This has been undertaken by Jakob and Dow (1946) who employed an electrically heated cylindrical specimen. Compared with the use of plane plates the cylindrical arrangement has the following advantages: A cylinder can be easily placed in the centre of an air jet and is free of the edge losses of a plate; for both reasons, air jets of moderate diameter can be used. Uniform heating is easier to provide, heat losses to the rear are easier to control, and noses of different shape and cylindrical starting sections can readily be used for studying the behaviour of the boundary layer of the fluid which is developing along the surface, first streamlined and then turbulent, and in which all resistance against heat transfer is concentrated. The experiments were performed with specimens of 1·3 inches diameter and 9 to 20 inches total length, the ratio of the heated length to the total length being varied from 0·4 to 0·9. Spherical, ellipsoidal, and conical nosepieces were used. The air velocity was varied from 10 to 150 ft./sec.

Figures 20, 21 and 22 show the general arrangement and details of the heating and supporting tubes. The heating coil consisted of nichrome wire wound on a stainless steel tube. Paper rings insulated the heating tube from the wooden starting piece or nosepiece and from the supporting steel tube; the junctions were smoothed by means of paraffin. Eight fine copper-constantan thermocouples for the determination of the surface temperature, t_s, were placed in four axial slots machined in the copper tube. Another thermocouple served to measure the temperature t_a of the airstream. Different secondary thermocouples in the wooden nosepieces and in the rear part of the heating-coil tube were used for heat-loss measurements. The air velocity, v_a, was measured by a Pitot tube.

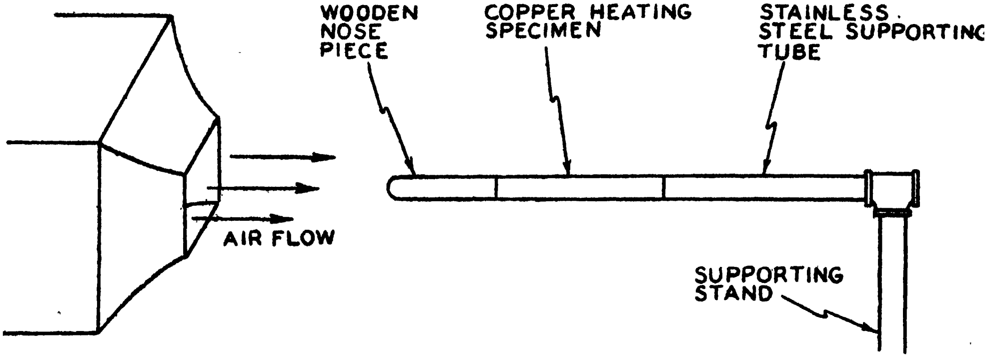


Figure 20. Arrangement for measurement of heat transfer in parallel flow.

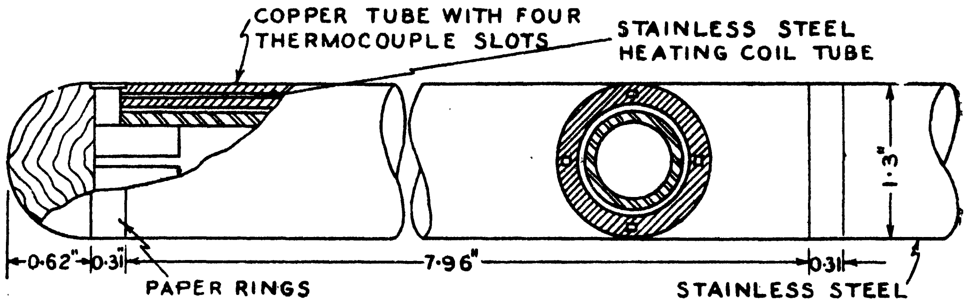

COPPER TUBE WITH FOUR THERMOCOUPLE SLOTS
STAINLESS STEEL HEATING COIL TUBE
1·3"
0·62"
0·31
7·96"
0·31
PAPER RINGS
STAINLESS STEEL

Figure 21. Heating specimen for parallel flow experiments.

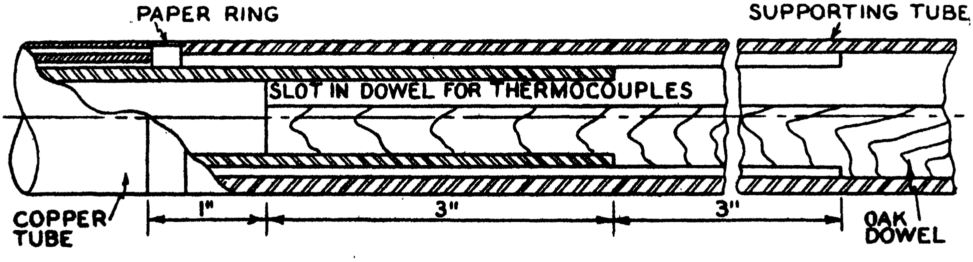


Figure 22. Supporting assembly for parallel flow experiments.

A hemispherical nosepiece is shown in figure 21. The six different starting pieces used in the experiments are described in table 3. In this table L_{st} is the hydrodynamic starting length of the specimen, defined as the ratio of the surface area of the starting piece to the perimeter of the heating cylinder. The total length, L_{tot}, is defined as the sum of L_{st} and the thermal or heating length, L_{th}, of 8 inches.

The experimentation consisted of velocity traverses, measurements of the input of the electrical heating coil, and thermoelectric measurements. The velocity traverses revealed that for a distance 2 feet downstream from the nozzle, the air velocity varied less than 1% within an 8-inch core. The rate of heat losses, q_j, including also radiation, was found to be 2·8 to 6·7% of the heat input, q_i. The mean coefficient of heat transfer by convection was found from the equation

$$h=\frac{q_i-q_j}{A(t_s-t_a)},$$

where A is the area of the heating surface.

Table 3. Starting pieces

Specification	L_{st} (ft.)	L_{tot} (ft.)	L_{st}/L_{tot}
Cylinder with hemispherical nose	1·026	1·693	0·606
Cylinder with hemispherical nose	0·689	1·356	0·508
Cylinder with hemispherical nose	0·354	1·021	0·347
Conical piece (4 in. long)	0·187	0·854	0·220
Ellipsoidal nose	0·092	0·759	0·122
Hemispherical nose	0·075	0·742	0·101

The experimental results were expressed in terms of Reynolds and Nusselt numbers, defined by $N_{Re}=v_aL_{tot}/\nu$ and $N_{Nu}=hL_{tot}/k$, where ν is the kinematic viscosity and k the thermal conductivity of the air. ν was taken at the temperature t_a, k at the temperature $(t_a+t_s)/2$.

By plotting N_{Nu} versus N_{Re} it was found that transition to turbulence started at $N_{Re}=60000$ to 200000 and was fully developed at $N_{Re}=250000$ to 600000.

For the range of laminar boundary layers the results could be represented by the equation

$$N_{Nu}=0{\cdot}590(N_{Re})^{0{\cdot}5}; \qquad ......(22)$$

for the range of fully established turbulence by

$$N_{Nu}=0{\cdot}0280(N_{Re})^{0.80}[1+0{\cdot}40(L_{st}/L_{tot})^{2.75}]. \qquad ......(23)$$

Figure 23 shows these relations and the transition from laminar to turbulent boundary layer.

Comparing these equations with those found theoretically for heat transfer in the flow parallel to plane plates, it is seen that equation (22) is in excellent agreement with the equation derived theoretically by Pohlhausen (1921) for heat transfer in the flow parallel to a plane plate, whereas the constant factor of equation (23) exceeds the one according to Latzko's (1921) derivation by 11%. In Jakob and Dow's paper it is shown that this is probably due to the surface curvature. It is further shown that Fage and Faulkner (1931) came to much higher values of the constant in equation (22), probably because starting conditions prevailed in their surfaces of only 0·333 to 1·27 cm. length. The only experiments in the turbulent range which can be compared with ours seem to be those of Juerges (1924) performed with a plane plate and yielding 15% higher values than we obtained. A blunt leading edge may have caused this increase in heat transfer. Eliás' (1929, 1930) values, which lie between Juerges' and our results, scatter considerably. A recent paper of Eckert and Drewitz (1940) shows that Pohlhausen's theory is valid up to twice the velocity of sound if the fluid temperature is replaced by a temperature impressed on the surface, due to adiabatic stopping and friction of the stream. The same is claimed for the turbulent boundary layer. It may be concluded that, when the equilibrium fluid temperature is replaced by the "impressed temperature" of the surface, the results of Jakob and Dow can also be approximately employed to much higher than the experimental velocities. Considering that our results in the turbulent range are almost between the theoretical ones of Latzko and the experimental ones of Juerges, it seems to be safe to use equation (23) tentatively for any convex smooth surface with arbitrary starting section where the radius of curvature in a plane perpendicular to the main flow direction is more than ½ inch.

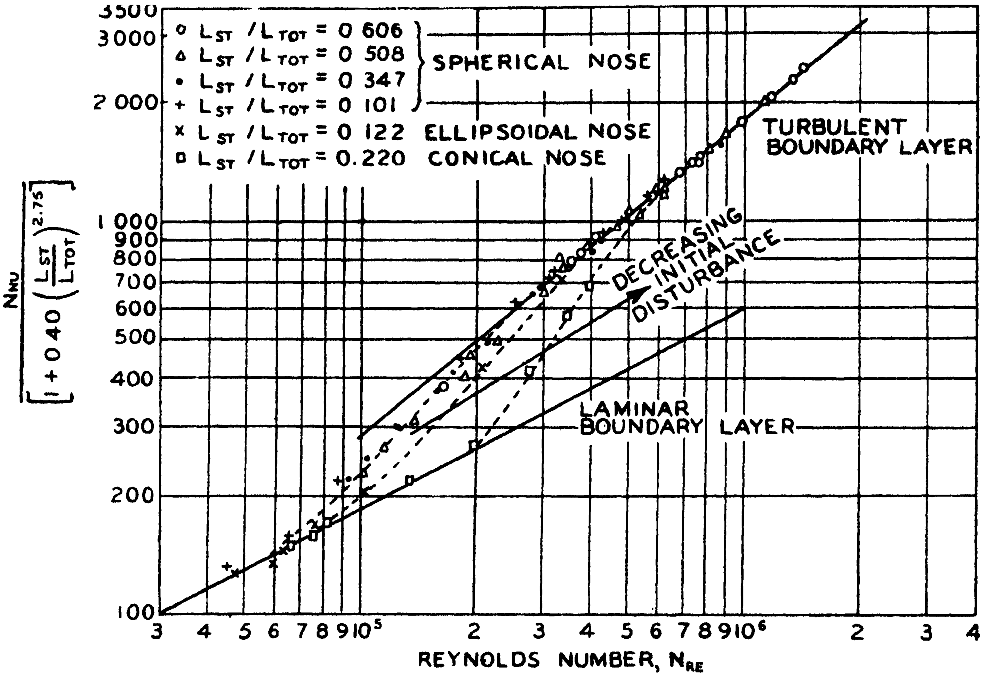


Figure 23. Heat transfer in parallel flow.

(d) *Forced heat convection in turbulent flow against a surface.*

Surprisingly little has been published regarding the heat transfer in the flow of air perpendicular or oblique to a large surface. There is the following statement of Reiher (1929):

"In blowing air perpendicularly to a surface, coefficients of heat transfer have been measured which, depending on the air velocity, were seven to eight times those determined by Nusselt and Juerges in flow parallel to the plate surface."

However, no experimental data or theoretical deductions have been published to substantiate this statement according to which the heat transfer close to the stagnation point would be unusually great. On the other hand, experiments by Rowley and Eckley (1931) at air velocities of less than 45ft./sec. indicated that the heat transfer of air impinging upon a surface perpendicularly is appreciably smaller than for parallel flow.

Since the heat transfer in the flow of air perpendicular or oblique to a surface plays an essential rôle in the formation, melting, and sublimation of ice on windshields and other parts of the outer surface of an airplane, experiments were sponsored by the U.S. Army Air Forces for the purpose of deciding between the above mentioned contradictory results, extending experience to higher velocities, furthering the understanding of the heat transfer in the flow of air against surfaces, and using the results for the calculations of ice formation and sublimation on wind-shields of airplanes. Two weeks before this lecture I presented a report on these experiments to the 6th International Congress for Applied Mechanics in Paris (Jakob and Kezios, 1946).

Owing to the practical purpose of the investigation one might have considered it as most promising to perform experiments under conditions of environment favourable to ice formation. However, quantitative experimentation imitating actual flight conditions would have been quite intricate. Fortunately, results obtained under conditions most convenient for laboratory tests can be converted to actual atmospheric conditions encountered in flight, by means of the theory of similarity. For this reason the experiments were performed with air approximately at standard atmospheric conditions and with surfaces at relatively high temperature (about 212° F.). The principle of similarity was then used to convert the results to conditions of flight at great altitude, i.e., low air pressure, low temperature, and considerably higher air velocities than were available in our laboratories. Finally, the theory of similarity between heat and mass transfer was used to calculate the amounts of ice that would be formed or sublimed on an exposed surface under conditions of flight.

The method of investigation was to expose the test plates to a homogeneous air-jet produced by a blower, or to bring them into the test section of a wind tunnel. The plates were heated from the rear by condensing steam and the heat transfer was determined by the rate of steam condensed (condensing-steam calorimeter). The temperature differences were measured by thermocouples. the flow velocities by Pitot tubes.

Figure 24 is a cross-section of our calorimeter No. 1. Its main parts are a heating plate A of 4 inches diameter and a guard-ring B, both of copper and chromium-plated, which are separated by air, except for the thin paper ring C; the main steam chamber D and the guard steam chamber E which prevents heat

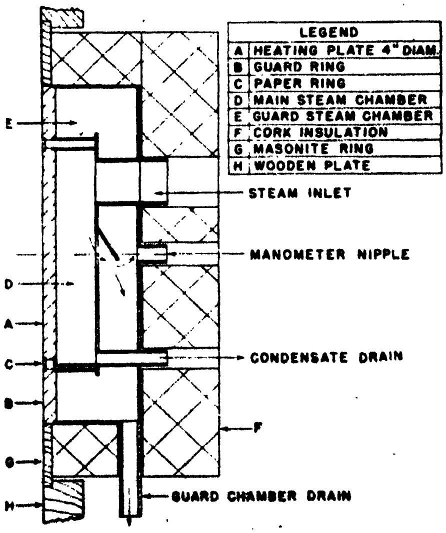


Figure 24. Calorimeter No. 1 for measurement of heat transfer in air flow against a plate.

loss from D; auxillary parts are the cork insulation F and the Masonite ring G which is inserted into the wooden plate H of rectangular shape, 34 inches (horizontal) × 37 inches (vertical).

Chamber D has its inlet at the top and its condensate outlet at the bottom; Chamber E is fed from the back of D. Four copper-constantin thermocouples were placed in holes drilled below the exposed surface of the calorimeters. Two more were attached to the back of the paper ring and served to determine the heat loss through this ring. A seventh thermocouple was used to measure the temperature of the incoming air.

Figure 25 is a schematic diagram of one of the experimental arrangements. The calorimeter was exposed to a free air stream delivered by a blower and could be turned around a horizontal axis to bring the surface into a position oblique to the air stream. Steam was formed in a small electrically heated boiler and dried by mechanical type steam separators.

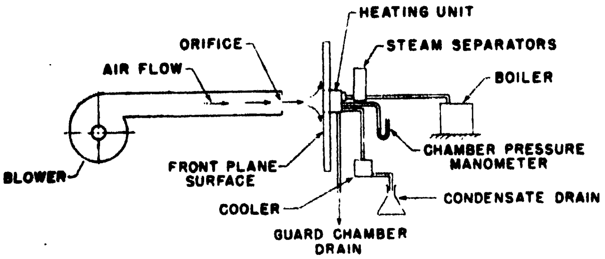


Figure 25. General arrangement for measurement of heat transfer in air flow against a plate.

Figures 26 and 27 are photos of the arrangement of our calorimeter No. 2, in the open section of the wind tunnel (throat dimensions 48 inches (horizontal) × 28 inches (vertical); distance between throat and diffusor 29½ inches). This calorimeter, whose main heating surface had a diameter of 2 inches, was inserted in a wooden plate of elliptical shape (axes 20·6 and 12 inches). Since the cross-sectional area of the free jet was 467 times that of the main heating surface, the conditions imposed approached those of the small stagnation area of a rather large surface in an infinitely wide air stream. Two steam separators which were connected in series and insulated by magnesia, and other apparatus, were placed in the wind shade on the rear of the wooden plate (see figure 27).

Experimentation consisted of measurements of air pressure and velocities, temperatures of air and calorimeter, steam pressure and condensate weight. The mean jet velocity as well as the velocity components parallel to the heat transferring surface were determined by traverses, the latter ones at $\frac{3}{16}$ inch distance from the plate surface.

Figure 28 is a sketch of the flow distribution over the exposed surface which is at an angle α from the vertical plane. For $\alpha = 0°$ the flow would have its stagnation point at the centre O of the plate for reasons of symmetry; at an angle $\alpha \lessgtr 0$ the stagnation point will be shifted, for instance, to S. The velocities on the surface were determined as follows (see figure 29):

First the stagnation point was found as that point where impact and static pressure tubes showed identical values. The air flow over the heating plate is limited between the lines SL and SR. The radii, SL, SC, and SR and the arcs 1-1, 2-2, and 3-3 intersect in 9 points. At these points the radial velocity (with S as centre) of the air was determined.

The scope of five series of experiments with a total of 61 runs is given in table 4.

Table 4. Experiments with air flowing against a heating surface

Series	Run No.	Calori-meter No.	Air jet Source	Air jet Outlet	Air jet Thickness (in.)	Distance between outlet and plate (in.)	Direction of impact	Diameter of heating plate (in.)
A	1–20	1	Blower	Orifice	7 (diam.)	20	Perpendicular	4
B	21–32	1	Blower	Orifice	7 (diam.)	10	Perpendicular	4
C	33–37	1	Blower	Orifice	10 (diam.)	10	Perpendicular	4
D	38–46	1	Blower	Orifice	10 (diam.)	10	Oblique (α = 30°)	4
E	47–61	2	Wind-tunnel	Throat	48 × 28	16	Perpendicular	2

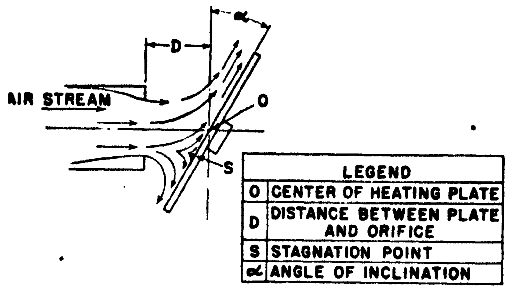

D
α
AIR STREAM
O
S
LEGEND
O CENTER OF HEATING PLATE
D DISTANCE BETWEEN PLATE AND ORIFICE
S STAGNATION POINT
α ANGLE OF INCLINATION

Figure 28. Sketch of air flow oblique to a plate.

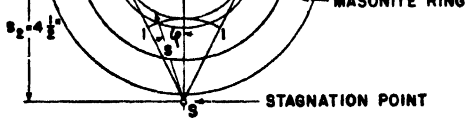


Figure 29. Range of air flow over inclined plate.

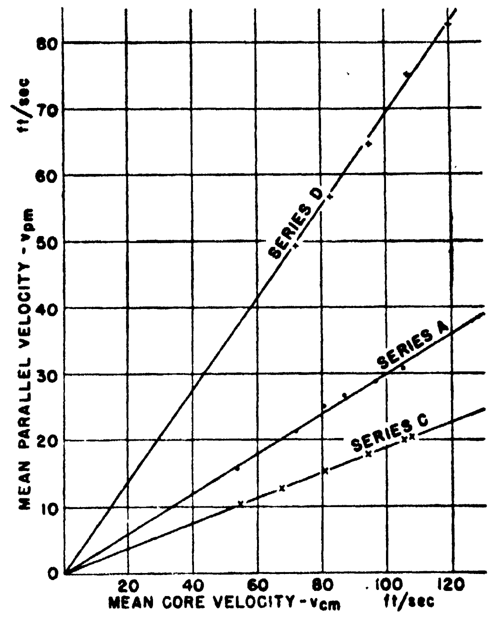


Figure 30. Parallel velocity versus core velocity in air flow against a plate.

The following definitions of velocities were used for representation and analysis of the experimental results:

The mean velocity of the core of the jet, v_{cm}, is the average velocity of that part of the jet which would hit the heating plate if the flow continued in the direction of the duct axis.

In Series A, B, C, and E the mean velocity parallel to the heating plate, v_{pm}, is identical with the mean radial velocity, v_{rm}, defined by the equation

$$v_{rm} = \frac{1}{\pi r_A^2} \int_{r=0}^{r=r_A} v_r \,.\, 2\pi r \,.\, dr,$$

where $r = r_A$ is the radius of the heating plate.

From figure 30 it is seen that, almost exactly,

$$v_{pm} = m \,.\, v_{cm}, \qquad \ldots\ldots(24)$$

where m is a constant for each series of tests ($m = 0{\cdot}300$ for Series A, $0{\cdot}191$ for C). This shows that the flow pattern in the experiments was not appreciably different from that of potential flow in which equation (24) should be strictly valid.

In Series D the stagnation point fell just outside of the Masonite ring (see figures 24 and 29), at 4½ inches from the centre of the plate, independent of the velocity. It was found that v_p was almost constant over each of the circular arcs 1-1, 2-2, and 3-3. Graphical interpolation and integration then led to

$$v_{pm} = \frac{1}{A}\int_A v_p \,.\, dA,$$

where dA is an element of the surface $A = \pi r_A^2$.

As in Series A and C, the relationship between the measured mean core and mean parallel velocities could be represented by a straight line through the zero point of the coordinates in figure 30. Hence equation (24) was also valid for Series D, the constant being $m = 0{\cdot}694$.

The heat transfer by convection was determined from the latent heat of the steam and the weight of condensate formed in the main calorimeter chamber with due consideration of conduction and radiation losses. In the runs of Series D the stagnation point fell outside the guard-ring. Hence, the air in the boundary layer was pre-heated in flowing over the guard-ring plate, whereas the mean film coefficient of heat transfer was defined under the assumption that unheated air meets the main heating surface. An approximate analysis showed that for this particular arrangement the influence of preheating was almost negligible.

In figure 31 the observed values of h_m are plotted against v_{cm} in logarithmic coordinates. The lines A, B, C, and D belong to the series denoted by these letters. The points of Series A are considerably scattered, particularly at low velocities. Conceivably, the relative small ratio of jet width to distance between orifice and test plate caused some instability in the stagnation region. The points of the other series are much better in line.

For $v_{cm} > 50$ ft./sec. the film coefficient could be represented by

$$h_m = N c_{mc}^{n}, \qquad \ldots\ldots(25)$$

with the constants, N and n, as given in table 5.

Table 5. Constants of equation (25)

Line	N	n
A	1·037	0·567
B	0·991	0·571
C	0·856	0·571
D	0·889	0·567

However, Series B may be as well represented with $N = 0{\cdot}991$ and $n = 0{\cdot}475$ in the whole range of v_{cm} from 14 to 124 ft./sec. (dotted line B′). This exponent is close to the theoretical value 0·5 for the streamline region in parallel flow.

The component of $n \approx 0{\cdot}57$, according to table 5, on the other hand, seems to indicate that in general a turbulent state was prevalent at $v_{cm} > 50$ ft./sec. This is almost the same exponent as has been found in the flow across a cylinder.

Regarding the differences of N, only one detail may be discussed here. Line D is close to C, though in Series D the plate was tilted, the stagnation point was outside the heating plate, and the mean parallel velocities were much greater ($v_{pm} = 69{\cdot}4$ ft./sec. compared to 19·1 ft./sec., both at $v_{cm} = 100$ ft./sec.). Since case D is closer to the conditions of a true parallel flow, a more pronounced impact effect in case C must have made good for the smaller parallel velocity.

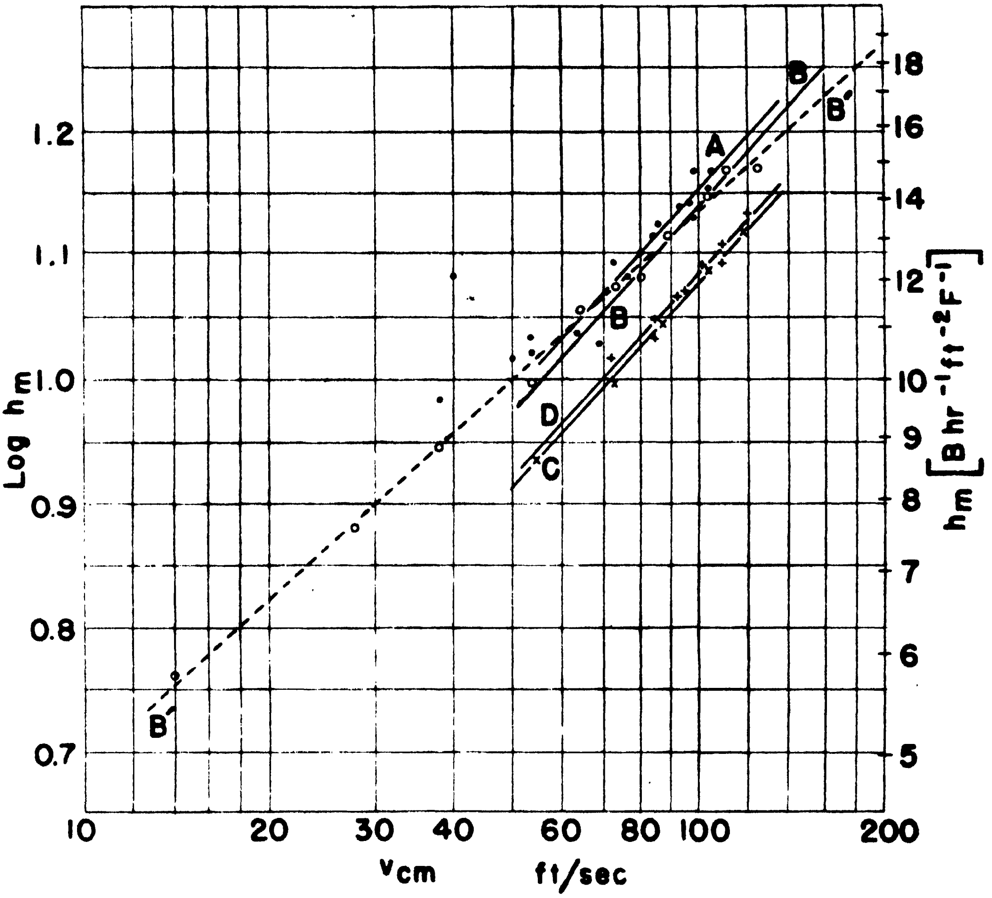


Figure 31. Heat transfer in air flow against a plate.

In the wind-tunnel tests with calorimeter No. 2, having a small heating surface, the heat transfer was so unstable that, notwithstanding all thinkable improvements of arrangement and frequent repetition of the tests, the points scattered by ±30%. However, the average in these tests, $h_m = 11$ for $v_{cm} = 90$ ft./sec., is in good agreement with the corresponding point in line C, indicating that an increase of the jet thickness from 10 to about 40 inches and of the distance from 10 to 16 inches did not appreciably change the heat transfer.

It may further be concluded that the coefficient of heat transfer close to the stagnation point, as observed in a circular area of 1 inch radius, is not much different from that observed in an area of 2 inches radius. No such singularity as mathematically might be expected seems to occur at the stagnation point, possibly just because the instability of flow masks any effect.

On the other hand, in strong contrast to Reiher's claim, the mean coefficient of heat transfer in the vicinity of the stagnation point remains much below that for parallel flow at the same jet velocity, and the impact compensates only partly for the considerable reduction of heat transfer in the stagnation region. At best, the impact effect may be compared with a strong starting disturbance in the flow parallel to a surface.

Several applications of the results to flying conditions have been given in our report to the U.S. Air Corps. It may be sufficient to show the procedure on one of these examples.

An airplane, rising from relatively low altitude through an atmosphere in which some ice is formed on its wind-shields, continues travelling at 30000 feet altitude in dry air of $-50°$ F. with the speed of 300 miles per hour. Assuming that the surface temperature is kept slightly below 32° F., calculate the rate of sublimation of ice into the air per unit area close to the stagnation point and the rate of heat to be delivered to the wind-shield from inside the plane and given up to the atmosphere under these conditions.

The calculation may be based on equation (25) with $N = 0{\cdot}87$ and $n = 0{\cdot}571$ holding for standard atmospheric conditions ($p = 760$ mm. Hg, $t_a = 68$ F.).

The equation is a special case of the general form

$$N_{\mathrm{Nu}} = C(N_{\mathrm{Re}})^n, \qquad \ldots\ldots(26)$$

where $N_{\mathrm{Nu}} = h_m L/k =$ the Nusselt number for a charateristic length L, $k =$ the thermal conductivity of the air, $N_{\mathrm{Re}} = vL\rho/\mu =$ the Reynolds number, $v =$ the flight velocity, $\rho =$ the density of the air, $\mu =$ the dynamic viscosity of the air.

Assuming the same characteristic length for the considered spot of the wind-shield in flight as in the laboratory experiment, $L = 1/6$ feet, N_{Re} is the same in both cases if

$$vL\rho/\mu = v_{cm0} L\rho_0/\mu_0 \quad \text{or} \quad N_{\mathrm{Re}} = (N_{\mathrm{Re}})_0, \qquad \ldots\ldots(27)$$

where subscript 0 refers to the laboratory test conditions. The air pressure at 30000 feet altitude is 226·1 mm. Hg, so that $\rho/\rho_0 = 0{\cdot}383$. Further, from physical tables $\mu/\mu_0 = 0{\cdot}825$. By substitution in equation (27), $v_{cm0} = 204$ ft./sec. Herewith, from equation (25), $h_{m0} = 18{\cdot}0$ B.hr.$^{-1}$ft.$^{-2}$F.$^{-1}$.

Since $N_{\mathrm{Re}} = (N_{\mathrm{Re}})_0$, equation (26) leads to $N_{\mathrm{Nu}} = (N_{\mathrm{Nu}})_0$. From physical tables, $k/k_0 = 0{\cdot}804$. Hence $h_m = kh_{m0}/k_0 = 34{\cdot}5$. Herewith the time rate of convective heat flow per unit area becomes

$$q_c'' = h_m(t_s - t_a) = 14{\cdot}5\ (32 + 50) = 1190 \text{ B.hr.}^{-1}\text{ft.}^{-2}.$$

According to the similarity of heat transfer and mass transfer, Nusselt (1930) has derived an equation for the case of small concentration of a diffusing vapour which, by combination with equation (26), becomes *

$$\frac{\dot{m}''}{q_c''} = \frac{\delta}{k}\left(\frac{\alpha}{\delta}\right)^n \frac{c_a - c_s}{t_a - t_s}, \qquad \ldots\ldots(28)$$

---

* Equation (28) was derived for diffusion and heat transfer not occurring in the same field. The more complicated formula for these processes taking place in the same field yields only a slightly different result.

where $\dot{m}''$ = the mass rate of sublimation for unit area in $\text{lb.}_{\text{mass}}\text{hr.}^{-1}\text{ft.}^{-2}$.
δ = the mechanical diffiusivity of water vapour into air in $\text{ft.}^2/\text{hr.}$
$\alpha = k/(\rho c_p)$ = the thermal diffusivity of air in $\text{ft.}^2/\text{hr.}$
ρ = the density of the air in $\text{lb.}_{\text{mass}}/\text{ft.}^3$
c_p = the specific heat of air at constant pressure in $\text{B.lb.}_{\text{mass}}{}^{-1}\text{F.}^{-1}$.
c = the concentration of water vapour in the air, in $\text{lb.}_{\text{mass}}/\text{cu. ft.}$

Subscripts a and s refer to bulk-air and surface conditions, respectively.

Assuming saturation of air with water vapour at the surface and entirely dry bulk air, from physical tables: $k = 0{\cdot}0119$; $\rho = 0{\cdot}0288$; $c_p = 0{\cdot}240$; $\alpha = 1{\cdot}72$; $c_a = 0$; $c_s = 303(10^{-6})$.

Further, from an equation of Mache (1874),

$$\delta = \frac{160}{p}\left(\frac{T_m}{492}\right)^{1.89} \text{ in cm}^2/\text{sec.}, \qquad \ldots\ldots(29)$$

where p is the air pressure in mm.Hg, and T_m is the absolute temperature of the air in degrees Rankine; $T_m = (T_a + T_s)/2$. This yields $\delta = 2{\cdot}325$ ft./^{2}hr. Hence, $\dot{m}''/q_e'' = 0{\cdot}000\,606\ \text{lb.}_{\text{mass}}/\text{B}$, $\dot{m}'' = 0{\cdot}785\ \text{lb.}_{\text{mass}}\text{hr.}^{-1}\text{ft.}^{-2}$.

The rate of heat flow due to sublimation per unit area is

$$q_{ig}'' = \dot{m}''\lambda_{ig}, \qquad \ldots\ldots(30)$$

where λ_{ig} = the heat of sublimation.

With $\lambda_{ig} = 1219{\cdot}1\ \text{B/lb.}_{\text{mass}}$ (from steam tables) one obtains $q_{ig}'' = 970\text{B.hr.}^{-1}\text{ft.}^{-2}$.

Though the temperature and concentration difference are large, only a small amount of ice, namely a layer of about 1/6 inch, will be sublimated in one hour of flight. The total rate of heat needed to keep the surface temperature at 32° F. and to sublimate this ice will be

$$q'' = q_c'' + q_{ig}'' = 1190 + 970 = 2160\ \text{B.hr.}^{-1}\text{ft.}^{-2}.$$

By moderately increasing the outer surface temperature, the ice could be melted. This would be preferable because only the relatively small melting heat (143 B./lb.) instead of the large heat of sublimation (1219 B./lb.) would have to be delivered from the inside of the plane and the water be wiped away by the air stream or mechanical devices.

It should be kept in mind that the above calculation is based on conditions close to the stagnation point. Since the heat transfer will be larger in the regions of parallel flow, a greater heat output will be needed in such regions in order to prevent freezing.

When I submitted this and similar calculations to the U.S. Air Corps I did not feel so confident, since the application of the equations of similarity between heat and mass flow to the present cases seemed not to be proved as yet by experiments. When, a few weeks ago, I was visiting the National Physical Laboratory, I learned, to my great satisfaction, that Drs. Griffiths and Powell had previously done and published appreciable work on evaporation and sublimation which during the war had not come to my attention. Concerning the method used by them, reference is made to their first paper (Powell and Griffiths, 1935). Powell (1940), in particular, has measured the evaporation of water from circular disks facing wind and expressed the measured values of the mass flow as a function of the 0·56 power of the Reynolds number, which is very close to the 0·57th power determined in our experiments on heat flow. Moreover, he (Powell, 1939, 1940) also measured the sublimation of ice on a sphere and came to an exponent 0·62. Considering that he found 0·59 for the sphere in evaporation experiments, which is more than 5% higher than for a circular disk, it may be assumed that 0·62 found for sublimation on a sphere would correspond to a 5% smaller value, that is, to 0·59, for sublimation on a disk, so that either his exponent 0·56 or 0·59 would have to be compared with our value 0·57 found in heat transfer experiments. I consider this very satisfactory agreement as a confirmation of Powell and Griffiths' experiments as well as of ours in the two different fields of observation.

§ 5. CONCLUSIONS

As mentioned in the introduction, I was not able to deal with general principles in this lecture, but only with some problems of heat transfer which occurred to me and had to be solved more or less exactly in one way or the other. Though many other methods of physics and mathematics have been employed in the field of heat transfer, the examples presented may have given you an idea of the kinds of procedure generally used in this branch of science. I also hope that you will have felt some satisfaction and stimulation due to the occupation with a variety of practical problems all of which can be reduced and are subordinated to a few general laws. In fact, recognizing possibilities of generalization in dealing with a special engineering problem not only raises the practical value of the work, but also causes a state of elation which is a sort of reward to those who take part in the scientists' mission to

" Seek the familiar law in chance's frightening wonder,
Seek the immovable pole in the phenomena's flight."

REFERENCES

American Standards Association, 1943. " American standard letter symbols for heat and thermodynamics, including heat flow " (published by *Amer. Soc. Mech. Engrs.*).

Davidson, W. F., Hardie, P. H., Humphreys, C. G. R., Markson, A. A., Mumford, A. R and Ravese, T., 1943. *Trans. Amer. Soc. Mech. Engrs.*, **65**, 553.

Eckert, E. and Drewitz, O., 1940. *Forsch. Geb. Ingwes.*, **11**, 116 (translated 1943 in U.S., N.A.C.A. Techn. Memor. No. 1045, Washington, D.C.).

Elenbaas, W., 1942. *Physica*, **9**, 1.

Éliás, Fr., 1929 and 1930. *Z. angew. Math. Mech.*, **9**, 434 and **10**, 1, respectively. Also 1930. *Abhandlungen Aerodynam. Inst. Techn. Hochschule Aachen*, Nos. 9, 10 (translated 1931 in U.S., N.A.C.A. Techn. Memor. No. 614, Washington, D.C.).

Fage, A. and Falkner, V. M., 1931. *Great Britain Advisory Committee for Aeronautics, Rep. and Memor.* No. 1408.

Griffiths, E. and Davis, A. H., 1922. *Special Report* No. 9 (London: Dept. Scient. and Industr. Res.).

Jakob, M., 1919. *Arch. Elektrotechn.*, **8**, 117; 1938. *Trans. Amer. Inst. Chem. Engrs.*, **34**, 173; 1939. *Ibid.*, **35**, 563; 1943 a. *Trans. Amer. Soc. Mech. Engrs.*, **65**, 581; 1943 b. *Ibid.*, **65**, 593; 1944. *Combustion*, **16**, No. 2, 49; 1946. *Trans. Amer. Soc. Mech. Engrs.*, **68**, 189; 1947. *Ibid.* (in printing).

Jakob, M. and Dow, W. M., 1946. *Trans. Amer. Soc. Mech. Engrs.*, **68**, 123.

Jakob, M. and Fritz, W., 1931. *Forsch. Geb. Ingwes.*, **2**, 435.

Jakob, M., and Hawkins G. A., 1942. *J. Appl. Phys.*, **13**, 246.

Jakob, M. and Kezios, St. P., 1946. *Minutes* of 6th Internat. Congr. for Applied Mechanics (in printing).

Jakob, M. and Linke, W., 1933. *Forsch. Geb. Ingwes.*, **4**, 75 ; 1935. *Phys. Z.*, **36**, 267.

Jakob, M. and Rees, K. A., 1941. *Trans. Amer. Inst. Chem. Engrs.*, **37**, 619.

JUERGES, W., 1924. *Beihefte zum Gesundheitsingenieur*, Reihe 1, Beiheft 19 (Germany: Muenchen and Berlin).
KING, W. J., 1932. *Mech. Eng.*, **54**, 347.
KOENIGSBERGER, J., 1903. *Ann. Phys., Lpz.*, **12**, 342.
LATZKO, H., 1921. *Z. angew. Math. Mech.*, **1**, 268 (translated 1944 in U.S., N.A.C.A. Techn. Memor. No. 1068, Washington, D.C.).
LORENZ, L., 1881. *Ann. Phys. (Chem.)*, **13**, 582.
MACHE, H., 1874. *Sitzungsber. k. Akad. Wiss. Wien (Math.-Naturwiss. Kl.)*, **1**, 385.
MULL, W. and REIHER, H., 1930. *Beihefte zum Gesundheits-Ingenieur*, Reihe 1, Heft 28 (Germany : Muenchen and Berlin).
NUSSELT, W., 1923. *Forsch.-Arb. Geb. Ingwes.*, Heft. 264 (Germany : Berlin) ; 1930. *Z. angew. Math. Mech.*, **10**, 105.
POHLHAUSEN, E., 1921. *Z. angew. Math. Mech.*, **1**, 115.
POWELL, R. W., 1939. *Proc. Brit. Assoc. Refrig.*, **36**, 1 ; 1940. *Trans. Inst. Chem. Engrs.*, **18**, 36.
POWELL, R. W. and GRIFFITHS, E., 1935. *Trans. Inst. Chem. Engrs.*, **13**, 175.
REID, W. T. and COREY, R. C., 1944. *Combustion*, **15**, No. 8, 30.
REIHER, W., 1929. In W. Wien and F. Harms, *Handb. Experimentalphysik*, **9**, part 1, 313 (Leipzig).
ROWLEY, F. B. and ECKLEY, W. A., 1931. *Amer. Soc. Heating and Ventilating Engrs.*, **3**, 870.
TOULOUKIAN, Y. S., HAWKINS, G. A. and JAKOB, M., 1947. *Trans. Amer. Soc. Mech. Engrs.* (in printing).

ION CONCENTRATIONS IN SPARK CHANNELS IN HYDROGEN

BY J. D. CRAGGS AND W. HOPWOOD,
Metropolitan-Vickers Electrical Co. Ltd.

MS. received 18 *October* 1946

ABSTRACT. Ion concentrations in hydrogen spark channels may, it is shown, be found by observing the Stark broadening of the Balmer lines. The line breadths are measured in two ways : by the normal techniques of photography or by plotting the line breadths with a photoelectric electron multiplier, amplifier and cathode-ray oscillograph. The advantages of the latter method is that the light emission/time relation can be studied. Representative oscillograms are shown in the paper.

Full account must be taken of the fine structure of the lines in assessing their true breadths, and Holtsmark's theoretical analysis of the Stark effect for inhomogeneous fields is used for that purpose.

§ 1. INTRODUCTION

THE physical properties of spark channels have not, in general, been accurately determined, largely because of the experimental difficulties involved and the uncertain and often erratic nature of the discharge. For the present purpose a spark channel is defined as the path of a spark discharge between two electrodes after complete bridging of the gap by a streamer and, more particularly, after conduction across the gap has persisted for $\nless 0{\cdot}25$ microsec. A gap 5 mm. in length would be bridged by a streamer in $\sim 3 \times 10^{-8}$ sec. or less (Loeb and Meek, 1940). The consecutive stages of avalanche, streamer and established channel are shown, for example, by Raether (1949) using Wilson-chamber techniques. Other papers dealing with spark channels in the present sense are, for example, by Flowers (1943) and Fucks and Bongartz (1943).

Modern techniques enable controllable spark sources to be used, especially with rapidly recurrent (50–1000/sec.) sparks at low currents ($\not>$ 500 amp.). The present experiments are a continuation of earlier work (Craggs and Meek, 1946) carried out with such techniques but are still to be considered as preliminary to the investigation of higher current discharges. The properties of short low-current sparks are of interest because of their importance in many aspects of the performance of electrical apparatus, and more particularly in spark-ignition problems and in spectroscopic analysis, etc., and also because their behaviour is probably in many ways reproduced in a slightly modified form in longer sparks or in those in which much higher currents are used. The latter are also of great technical interest.

The main objective in the present work was to investigate the Stark broadening of the Balmer lines for hydrogen sparks and so to deduce the ion concentrations in the channels. Early work was carried out by Lawrence and Dunnington (1930) and by Finkelnburg (1931). Qualitative observations using spark sources whose characteristics were largely unknown were made, e.g. by Merton and Hulburt.

The most recent work is that of Finkelnburg, whose conclusions are ill defined and whose methods of analysis (different from that described in the present paper) of line broadening are very inaccurate. The results of this work are discussed in §6.

§2. DETAILS OF APPARATUS

The sparks were passed between pointed tungsten electrodes about 5 mm. apart in hydrogen at pressures 10 cm. Hg above the atmospheric value. The tungsten electrodes were mounted in collars attached in turn to electrodes sealed into the two halves of a demountable Pyrex bulb some 150 c.c. in volume. Since some photographs were required of the u.v. emission from the spark, all measurements were made through a quartz window waxed on to a tubulation in the side of the experimental tube.

The circuit used for supplying square-voltage pulses was that devised originally in this laboratory by Mr. M. E. Haine and Professor J. M. Meek for use in radiolocation modulators, and is described in some detail by Craggs, Haine and Meek (1946). The circuit is shown here in figure 1. L is the choke through which the artificial line M is charged from a high-voltage D.C. supply, provided by a half-wave rectifier set. It is shown (Craggs, Haine and Meek, *loc. cit.*) that it is advantageous to use a choke whose inductance is large compared with the value necessary to give resonance with the line capacity at half the required resonant frequency. The frequency can then be adjusted by alteration of the frequency of the incoming trip-pulse applied at T to the trigger electrode of the special three-electrode gap (Trigatron) shown in figure 1, developed and described by Craggs, Haine and Meek. In steady-state conditions the voltage on the line always builds up to nearly twice the D.C. charging voltage. The spark-gap current and voltage, or

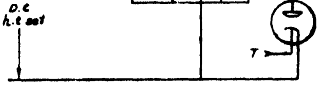


Figure 1.
H.T. source for spark excitation.

light output (figure 14), can be monitored by an oscillograph tripped synchronously from the circuit used to fire the Trigatron.

If it is required to study sparks in a two-electrode gap, such a gap may be included in the circuit of figure 1 between the artificial line and the Trigatron. The use of the resistive matching load R, necessary to give non-oscillatory square voltage pulses with the line, is not affected by the inclusion of the test gap, except possibly during the first 0·5 microsec., when the spark-gap impedance falls from infinity to ~2 ohms (hydrogen). The pulse time, for a constant peak current, was varied from 1 to 4 microsec. by inclusion of several artificial lines, all of constant impedance (about 80 ohms) in cascade. For the 10- and 20-microsec. pulses it was more convenient to use special lines of higher impedance and smaller bulk. The arrangement of two gaps was used in the earlier work of Craggs and Meek (1946), although the circuit was not there described in detail.

The D.C. supply voltage (figure 1) was measured with a calibrated high-resistance voltmeter and the test-gap voltage may be derived from that value by calibration of the charging circuit or by direct measurement with a sphere gap or calibrated oscillograph and potential divider.

The controlled sparks produced with the above circuit were observed with a spectrometer and also with a system comprising a photoelectric electron-multiplier, amplifier and cathode-ray oscillograph.

The spectrometer (small Hilger constant-deviation model) gave a spectrum, from 4000 to about 6700 A., some 4·5 cm. in length. In order to ensure uniform illumination of the spectrometer slit, and thus of the neutral step wedge (placed immediately before it) necessary to provide plate calibrations, a two-lens collimating system was used (figure 2). This arrangement also minimizes undesirable reflections in the collimator tube of the spectrometer. Figure 2 shows that an image of the spark was formed by L_1 on L_2 and an image of L_1 was projected by L_2 in the plane of the slit. The adoption of such a system is essential in work of this kind where the path of the spark varies slightly in position for successive discharges. A photograph of the slit, with step-wedge removed, was taken on each plate in order to ensure uniform illumination of the slit in all the analysed records.

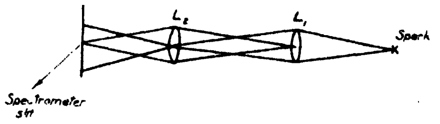


Figure 2.
Optical system for slit illumination.

The line profiles and step-wedge marks on the plates were measured with a Hilger microphotometer in the University of Manchester. One of us (W. H.) in collaboration with W. K. Donaldson modified the instrument to a self-recorder, and this work will be described in a separate publication. The microphotometer sensitivity was such that a net deflection of 15 cm. was obtained for a change from zero to infinite plate density. The magnification could be made either 140, 70 or 35 to 1 by choice of suitable gears in the automatic drive, and vertical lines were flashed on to the records at intervals of 0·05 mm. along the photographic plate.

After many preliminary experiments, Ilford S.G. Panchromatic plates were adopted since they were found to have the most suitable spectral response for the particular Balmer series decrements obtained with the sparks. Each plate was carefully rubbed with cotton-wool during development to avoid spurious local variations in plate density, and the results showed clearly that the treatment was adequate.

The electron multiplier system in its earlier form was similar to that described by Craggs and Meek (1946). In order to improve linearity, a lower load resistance and a smaller multiplier output current were used for the present work, in conjunction with a VT60A amplifier valve to feed the oscillograph. The amplifier valve had its frequency response improved by the use of the circuit of figure 3 (see Brainerd *et al.*, 1942).

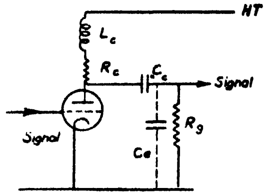


Figure 3.
Correcter circuit for H.F. amplifier.

A characteristic frequency f_e is defined by $2\pi f_e = 1/R_c C_e$ (C_e = effective capacitance shunting the coupling circuit). For $C_e = 20$ pf. and $R_c = 2200\,\omega, f_e = 3\cdot7 \times 10^6$ c./s.

A quantity D is defined by

$$D = \frac{L_c}{C_e R_c^2}.$$

The relative stage gain is plotted as a function of frequency f (Brainerd *et al.*, *loc. cit.*) and with D as a parameter. The following data are relevant:

Frequency f	D	Relative stage gain
f_e	0	0·7
	0·4	0·92
	0·5	1·0
	0·6	1·08
$1\cdot4 f_e$	0	0·58
	0·4	0·8
	0·5	0·88
	0·6	0·94

A value of $D \sim 0\cdot5$ is therefore desirable. In practice the value of L_c is best found by trial, using a square pulse generator to excite the amplifier. The results showed that only with the 1-microsec. pulses would an error arise, and even in that case the error would be negligible. The frequency response at 1 Mc./sec. was finally about 90%.

The usual load resistance for the VT60A was 15,000 ohms (which gives a time constant of 0·3 microsec. with the oscillograph input capacitance of 20 pfs.). Experiments were performed with the load varied from 15,000 to 3000 ohms and the effect on the spark light/time oscillograms was barely detectable. This system was used in preliminary work with new multipler tubes to confirm with greater accuracy some of the earlier experiments of Craggs and Meek (1946).

For the new experiments on plotting Stark profiles of the separate hydrogen lines it was necessary to use higher amplification. A three-stage h.f. amplifier, using SP41 valves with 2000 ω anode resistances and choke correction for improvement of frequency response, was then used in conjunction with the multiplier, and the VT60A valve was retained to provide a sufficiently great voltage swing for the oscillograph deflector plates. The frequency response was of the order of that for the VT60A alone, and careful tests were again carried out to check the frequency response. The linearity of response of the multiplier and amplifier system was checked with the spark in operation by interposition in turn of a number of calibrated gauzes between the light source (spark) and the multiplier.

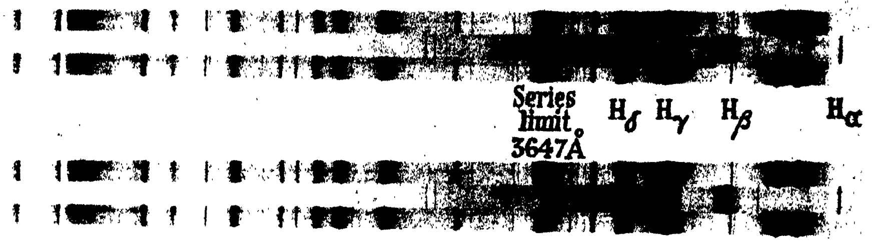


Figure 4. Spectra of hydrogen sparks (centre) with copper electrodes, and mercury vapour for comparison.

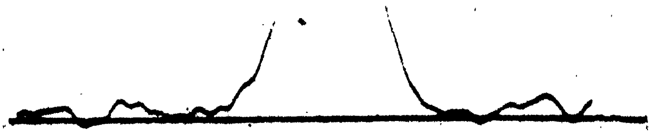


Figure 5. Microphotometer tracing of H_α for 1-microsec. current pulse.

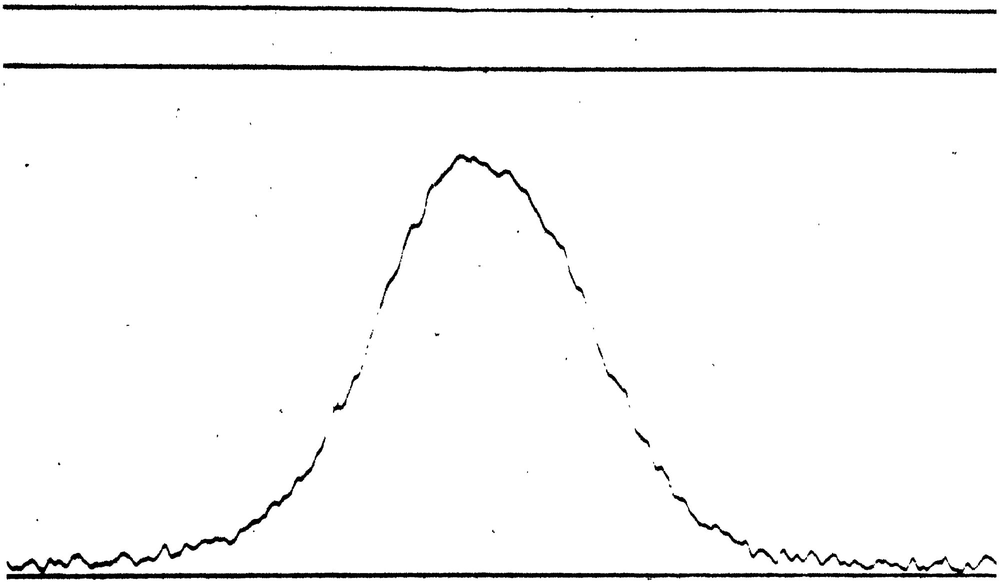
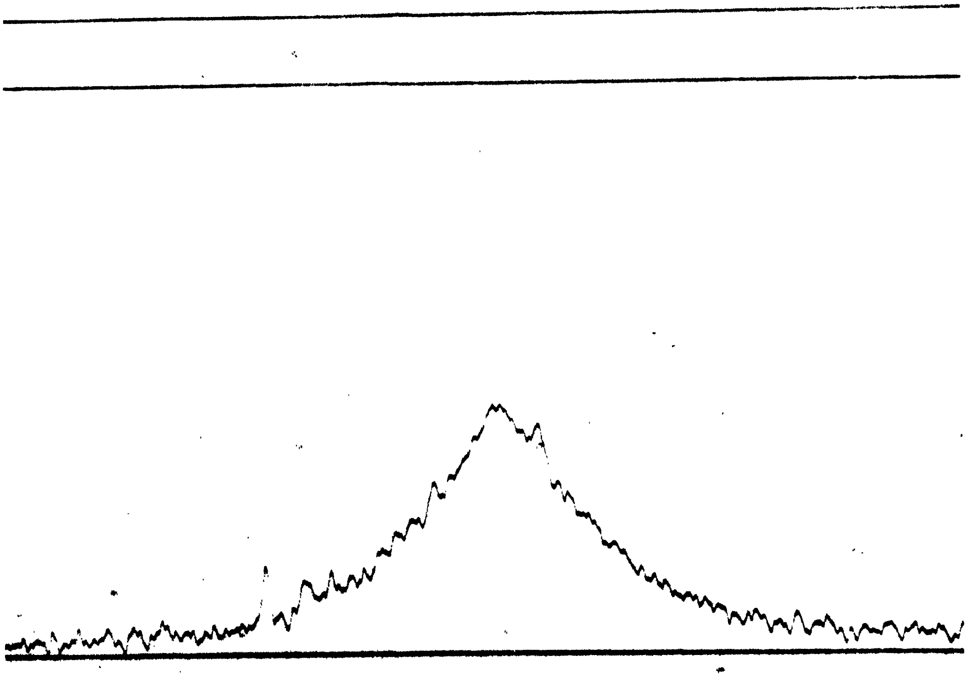


Figure 6. Microphotometer tracing of H_β for 1-microsec. current pulse.



It is of interest to note that the complete system gave an amplification of 10^9–10^{10} times, i.e. $\times 1{\cdot}7 \times 10^6$ with the multiplier, about × 300 with the three-stage amplifier and about × 10 with the VT60A. Considerable trouble with pick-up from the spark supply circuits was experienced, and even the final records (e.g. figure 14) show a small dip at the beginning of the oscillograms, where its presence is not important. If the multiplier was irradiated with feeble steady light, the random fluctuations in the photo-current were noticeable on the oscillograph screen, but with the stronger light from the sparks such effects were negligible. It was, however, extremely important to use a light-tight box for the multiplier.

§ 3. EXPERIMENTAL RESULTS WITH THE PHOTOGRAPHIC TECHNIQUE

Every plate was calibrated with the Hilger step-wedge, using microphotometer measurements taken at the peak of each broadened line. The general appearance of the H_α, H_β and H_γ lines is shown in the spectrogram of figure 4, taken with a quartz prism spectrograph.

Experiments were made with pulse lengths of nominally 1, 4 and 10 microsec. with respective peak currents of 120, 120 and 30 amp. Figure 14(9) shows a typical 10-microsec. pulse. The 1- and 4-microsec. pulses were used in order to determine, if possible, the changes in ion concentration during the afterglow period which, being the same for both pulses and the same peak current, constitutes a greater fraction of the total time of light emission with the shorter, i.e. 1 microsec., discharge. Further reference to afterglows are made in § 6.

Representative microphotometer tracings, using 1-microsec. pulses of peak current 120 amp., are given in figures 5, 6 and 7 for H_α, H_β and H_γ respectively. The fact that the shape of H_α was not distorted by the time lag of the galvanometer system in the microphotometer was confirmed by taking a slow manual plot of H_α.

The microphotometer tracings were corrected, by the use of plate response curves, and typical examples of such corrected tracings are given in figures 8 to 13. A complete set of profiles for H_α, H_β and H_γ for the pulse lengths 1 and 10 microsec. (pulses of figure 14) is reproduced here in order that, if desired, methods other than that described below for the estimation of ion concentration may be tested. The curves for the 4-microsec. pulses are the same as for the 1-microsec. pulses. One unit on the wave number scale for H_α, H_β and H_γ corresponds respectively to 31·2, 40 and 65 cm.^{-1}.

The analysis of these results is given in detail below (§ 5).

§ 4. EXPERIMENTAL RESULTS WITH THE ELECTRON-MULTIPLIER TECHNIQUE

In order to take plots of the broadened lines with the electron multiplier, the constant deviation Hilger spectrometer was fitted with a telescope adaptor using a standard slit taken from another spectrometer of the same type. The telescope-slit assembly was made to project into the metal box containing the multiplier, and stray light was thus eliminated (see § 2). It was essential, in order to avoid obtaining distorted line profiles, to focus the collimator slit on to the exit slit.

The colour response of the multiplier was measured in a subsidiary series of experiments, using the above spectrometer as a monochromator with a tungsten

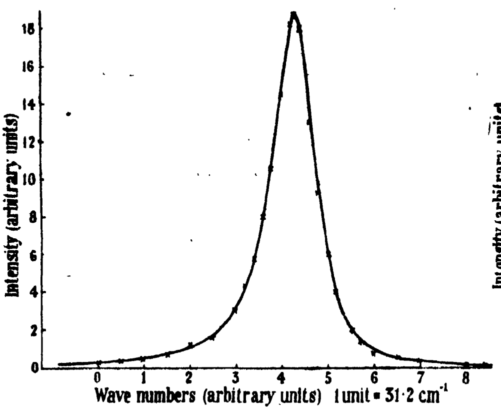


Figure 8. Profile of H_α with 1-microsec. pulse.

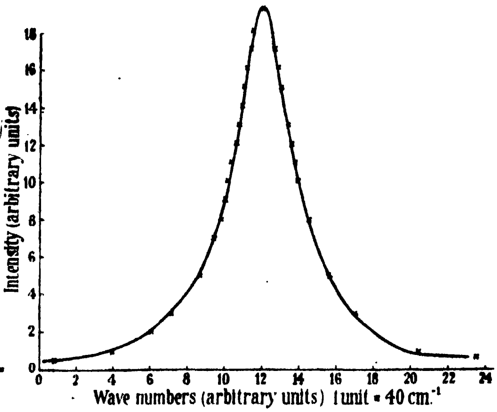


Figure 9. Profile of $H\beta$ with 1-microsec. pulse. Full line shows photographic results, crosses show multiplier results. The curves are scaled to the same size at the peaks and at one other point.

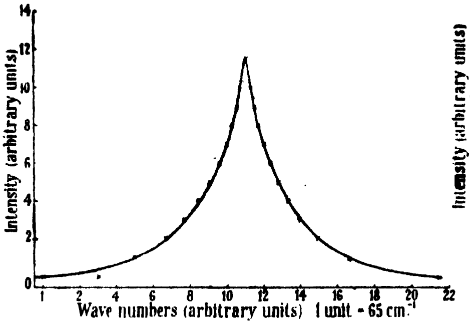


Figure 10. Profile of $H\gamma$ with 1-microsec. pulse. Full line shows photographic results ; crosses show electron multiplier results when the curves are scaled to the same size at the peak and at one other point.

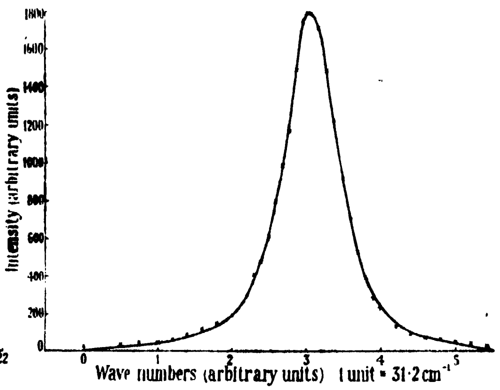


Figure 11. Profile of $H\alpha$ with 10-microsec. pulse.

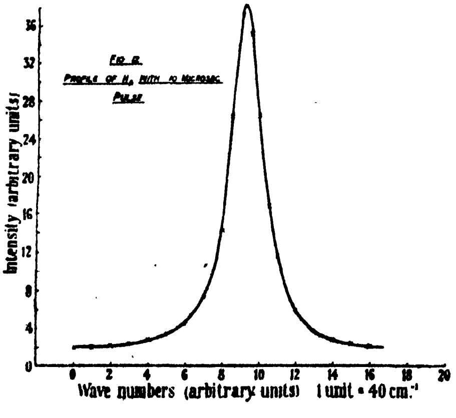


Figure 12. Profile of H_β with 10-microsec. pulse.

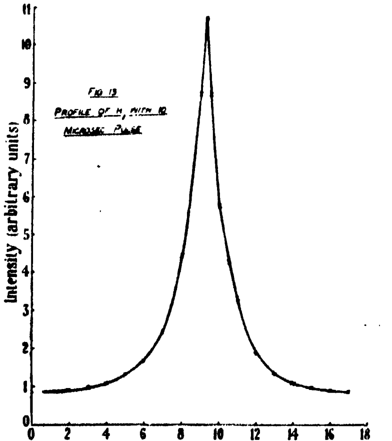


Figure 13. Profile of $H\gamma$ with 10-microsec. pulse.

filament lamp, running at known temperature, as source. The emission characteristics of such lamps are known and comprehensive data have been published by Forsythe and Adams (1945). The probably negligible but unknown differential absorption in the glass system of the spectrometer and external illumination system (figure 2) were the same since identical optical systems were used for the sparks and for the standard lamp and so did not introduce errors. The colour response of the multiplier, corrected for spectrometer dispersion, was used to check that the very wide line profiles of H_β and H_γ were not spuriously distorted by such colour response. The red response of the multiplier is extremely poor, and reliable profiles for H_α could not be obtained since it was not possible to reduce the telescope slit (normally about 50 microns wide) for the narrow H_α profiles. This difficulty did not arise with H_β and H_γ since the colour response was good and the line-breadths were such that a 50-micron slit was sufficiently narrow.

The fully corrected profiles for H_β and H_γ with 1-microsec. pulses and a peak current of 120 amp. are shown respectively in figures 9 and 10, and analysis of the data is given in § 5. The multiplier profiles in figures 9 and 10 are fitted to the photographic profiles at the peaks and at one other point, although (see tables 3 and 4) the multiplier profiles are slightly wider and thus give higher ion concentrations. The importance of figures 9 and 10 is that, after scaling, the photographic and multiplier curves should be found to be identical. This fact is discussed in § 6. The line profiles taken with the electron multiplier for 1- and 4-microsec. pulses were found to be identical.

Before proceeding to a discussion of the above results it is of interest to show spark-light emission as a function of time for the different Balmer lines. The records are similar to those shown for total (polychromatic) spark radiation by Craggs and Meek (1946), but are taken with the greatly improved techniques described above in §§ 2 and 3. The selected oscillograms are conveniently described in tabular form (table 1).

Table 1

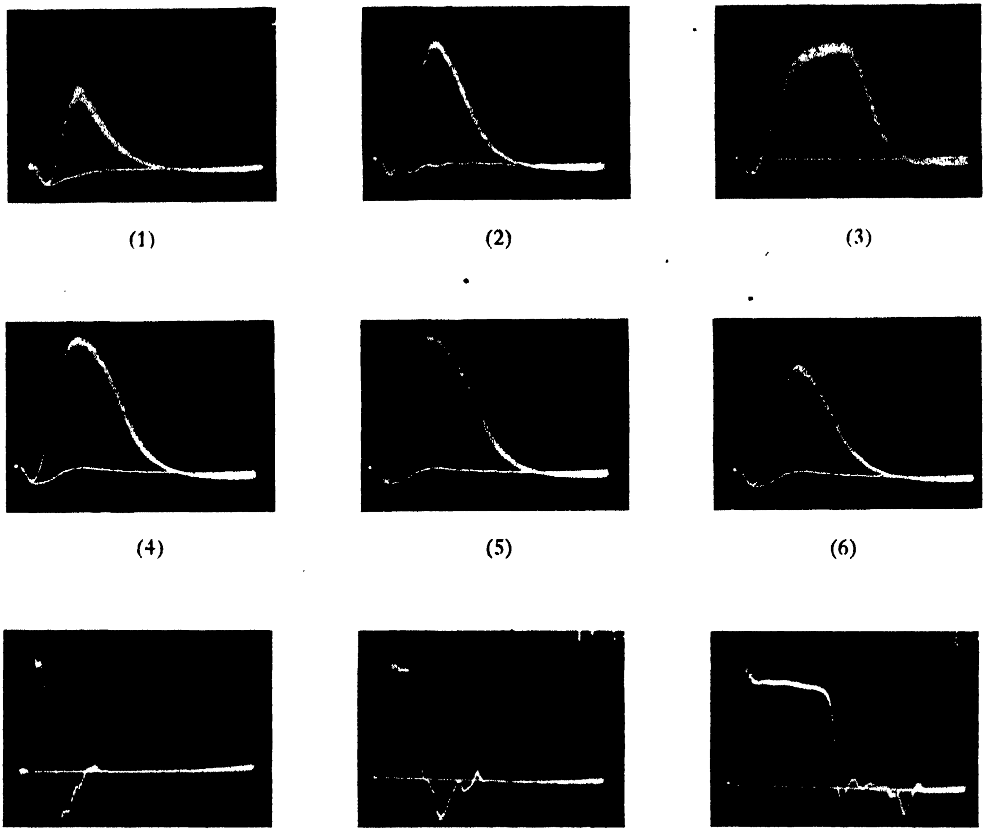
Light/time diagrams. Figure No.	Spectral line	Current wave		
		Nominal deviation (microsec.)	Figure No.	Peak current (amp.)
14·1	H_β	1	14·7	120
14·2	H_β	2	14·8	120
14·3	H_β	10	14·9	30
14·4	H_α	2	14·8	120
14·5	H_β	2	14·8	120
14·6	H_γ	2	14·8	120

It is noticeable that the shapes of the light/time diagrams for H_α, H_β and H_γ are the same within the close limits of observational error. It is noticeable that the smaller diagrams (e.g. figure 14(2)) are sharper than the larger ones (e.g. figure 14(5)), due to the slight non-linearity of the multiplier and amplifier system (see § 2). Hence the Balmer decrements, i.e. the intensity ratio of H_α to H_β and H_β to H_γ, do not change during the time of discharge, and this means in turn that the mode of excitation responsible for light emission is also constant for that period. The investigation of Balmer decrements will be discussed, it is hoped, in another paper. The afterglow discussed by Craggs and Meek (1946) is again apparent on comparing, for example, the duration of current and light in figures 14(1) and 14(7), and it is hoped in later experiments to determine Stark profiles for the broadened Balmer lines as a function of time and so including measurements taken during and after the flow of current. In § 5 the interpretation of results taken with different pulse lengths is given. It is clear that for afterglows of equal duration, and for the same peak currents, the effect of such an afterglow would be more noticeable for shorter pulses when time averages of light-output were determined. The afterglow shown in e.g. figures 14(1) and 14(2) is artificially long because of the slight mismatching of the artificial line as shown by the current pulses of figures 14(7) and 14(8). For that reason, the record of figures 14(3) and 14(9) and those given by Craggs and Meek (1946) are more satisfactory for measurements of afterglow, with which the present paper is not primarily concerned. It is further proposed that this work, now being continued for higher current discharges, where detectable afterglows are longer, will be extended, in collaboration with Professor J. M. Meek, to include experiments made also with different techniques.

§ 5. ANALYSIS OF EXPERIMENTAL RESULTS

It is well known that the Balmer lines are split into several discrete components if the radiating atoms are subjected to a uniform and uni-directional electric field. The Stark patterns are different according to the direction of observation, i.e. whether the latter is parallel or perpendicular to the field. In the case of inhomogeneous spherically symmetrical fields, the two sets of discrete components are merged into a continuous pattern, and it is this form of the Stark effect that obtains in spark and other discharges where the operative field is that due to the mutual actions of ions. Observations of the Stark effect in such conditions may thus be used to determine ion concentrations, which is the object of the present work. As mentioned in the Introduction, § 1, Merton (1915) noticed the peculiar form of the Balmer lines in such inhomogeneous fields but made no quantitative measurements. Holtsmark (1919) formulated a detailed theory of the Stark effect for such conditions, and even gave an approximate calculation of the ion concentration for an arc discharge in a gas containing lithium vapour voltatilized from the arc electrodes, using only the total width of the Stark pattern, ignoring fine structure, and comparing it with the separation of the outermost components of the lithium Stark pattern for uniform fields of known value. Holtsmark's analysis has been quoted by many other writers (e.g. de Groot, 1931; Weisskopff, 1933; and Margenau and Watson, 1936), none of whom applied the theory in detail to any particular practical case, although de Groot made approximate calculations, again ignoring the effects of fine structure, for a hydrogen arc. Hulbert (1923, 1924a and b, and 1926) made an instructive independent attempt (1923) to work out inhomogeneous field Stark profiles but admitted its approximate nature (1923, 1924a) and applied it to certain ill-defined practical cases (1924a and 1926). Holtsmark's fuller analysis seems preferable and is used in the present case, as very recent and exact work by Spitzer (1939a and b) accepts the validity

PROC. PHYS. SOC. VOL. 59, PT. 5 (J. D. CRAGGS AND W. HOPWOOD)



(1) (2) (3)

(4) (5) (6)

(7) (8) (9)

Figure 14. Oscillographic records of light-output time and current time for hydrogen sparks.

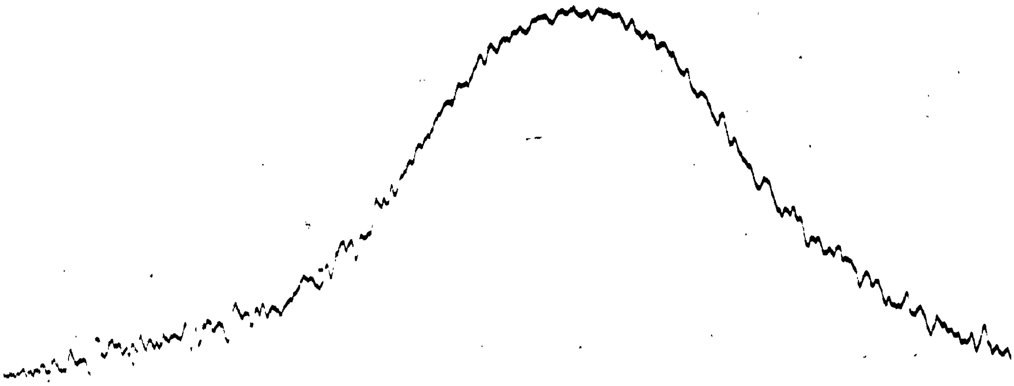


Figure 21. Microphotometer plot of spark channel for 1-microsec. current pulse.

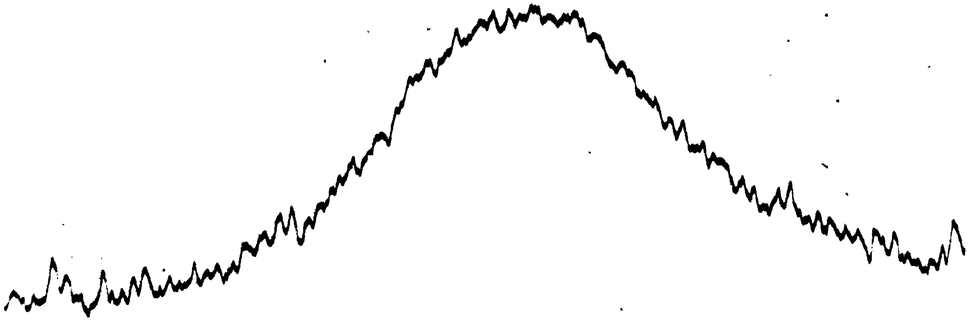


Figure 22. Microphotometer plot of spark channel for 4-microsec. current pulse.

of his treatment. As mentioned above, the profiles of H_α, H_β and H_γ are given in full (figures 8–13) in order that further formulae may be applied. Verweij (1936) had computed Holtsmark's probability function $W(F)$, which is explained below, and thus built up the Stark profiles for inhomogeneous fields in a manner similar to that described here. Verweij's results apply to stellar conditions and to absorption lines.

Holt mark's theory (1919) of the Stark effect in an inhomogeneous field leads to the equations

$$W(F)dF = \frac{4}{3\pi}\beta^2 d\beta[1 - 0{\cdot}4628\beta^2 + 0{\cdot}1227\beta^4 - 0{\cdot}02325\beta^6 \ldots.] \qquad \ldots\ldots(1)$$

and

$$W(F)dF = \frac{d\beta}{\pi\beta^{5/2}} 2{\cdot}350\left[1 + \frac{5{\cdot}106}{\beta^{3/2}} - \frac{7{\cdot}4375}{\beta^3} + \ldots.\right], \qquad \ldots\ldots(2)$$

where $W(F)dF$ is the probability of an atom being subjected to an electric field lying between F, $(F+dF)$ and

$$\beta = F/F_n, \qquad \ldots\ldots(3)$$

where F_n is an effective mean or normal field strength given by

$$F_n = C_1 e N^{2/3} \qquad \ldots\ldots(4)$$

for ionic fields, where C_1 is a constant calculated by Holtsmark (see also Weisskopff, 1933; Verweij, 1936) as 2·10, giving $C = 2{\cdot}61$, e is the electronic charge, and N the ion concentration (ions/c.c.). Other expressions similar to equation (4) hold for dipole and quadrupole fields, both of which may be ignored in the case of hydrogen (Margenau and Watson, 1936). The above equations evaluate the field acting on a particular radiating atom by virtue of the surrounding atoms, located at varying distances according to a classical distribution formula and so contributing individually in different amounts to the total field at the radiator. The complete Holtsmark profile for a line showing fine structure in a homogeneous field consists of a summation of intensity/wave-length patterns with one pattern for each component of the line. The maximum width of the line, for the case of inhomogeneous fields giving a normal field strength F_n, is the separation of the outermost components for a homogeneous field equal in magnitude to F, where $F = F_n$ for the first-order Stark effect. The second-order effect, in which the separation of a given component varies as (field)^2, is negligible for the fields considered here, as can be shown by a consideration of data for the second-order effect (Gebauer and von Traubenberg, 1930, and White, 1934). The second-order effect is only $\sim$10% of the first-order effect, in wave number separations, for $F = 10^{11}$ volts/cm.

Equations (1) and (2) apply respectively for $0 < \beta < 1{\cdot}7$ and $1{\cdot}7 < \beta < \infty$ respectively, and are plotted separately and added together in figure 15. For the computation of the resultant shape of a complex line, such as H_α, H_β or H_γ, the curve of figure 15 is applied to each component of the line and the different curves are then superimposed.

It is assumed that only the first-order Stark effect is operative, i.e. for any component

$$\delta\nu = CF, \qquad \ldots\ldots(5)$$

where $\delta\nu$ is the separation (cm.$^{-1}$) of the component from the undisturbed line in a homogeneous field F. C is a constant. The highest components disappear first as the field is increased, owing to level perturbations (Finkelnburg, 1931, and references there cited), and Finkelnburg used this fact and the available data to estimate inter-ionic fields in hydrogen sparks. The method is very inaccurate and uncertain, and Finkelnburg apparently made no attempt to carry out the full Holtsmark analysis (given here,) which uses Stark-effect data in a different manner. The level smearing is negligible for $F_n \sim 100$ kv./cm., which is the value obtained in the present work. The values of C were then found for all important components (table 2) of the different lines, using data given by Minkowski (1929) for $F = 104$ kv./cm. The weakest components were ignored.

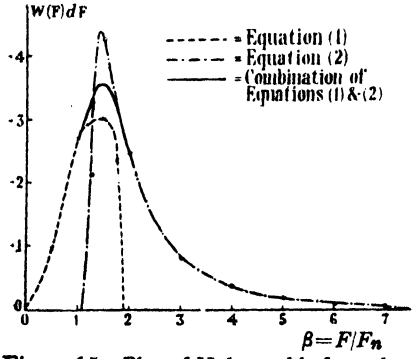


Figure 15. Plot of Holtsmark's formulae.

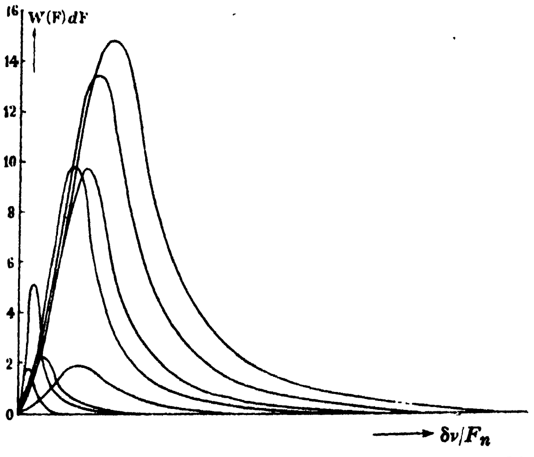


Figure 16. Holtsmark plots for components of H_γ.

Table 2

Line	$n's' - ns$ (quantum number term)	$\Delta\nu$ (cm.$^{-1}$)	C (cm.$^{-1}$/kv./cm.)	Relative intensity
H_α p components	2	14·4	0·138	729
	3	20·4	0·196	2,304
	4	26·6	0·256	1,681
H_α s components	0	0	0	5,290
	1	6·0	0·058	1,936
H_β p components	6	42·3	0·407	81
	8	55·7	0·536	384
	10	69·3	0·666	361
H_β s components	2	14·4	0·1385	72
	4	27·9	0·268	456
	6	41·0	0·394	294
H_γ p components	2	14·3	0·137	15,625
	5	35·0	0·336	19,200
	12	84·3	0·810	16,641
	15	105·0	1·01	115,200
	18	127·0	1·22	131,769
H_γ s components	0	0	0	141,650
	3	20·3	0·195	46,128
	10	70·7	0·680	88,050
	13	91·0	0·875	83,232

From equations (3) and (5)

$$\beta \times \text{constant} = \delta\nu/F_n, \qquad \ldots\ldots(6)$$

and the plot of $W(F)dF$ against β may be converted to a plot of $W(F)dF$ against $\delta\nu/F_n$, so that each fine-structure component is represented by a separate curve, the area under which is given by

$$A = \int_0^\infty W(F)dF\,.\,d\left(\frac{\delta\nu}{F_n}\right). \qquad \ldots\ldots(7)$$

If the areas obtained for the curves drawn for the different components are arranged by scaling to be proportional to their respective relative intensities (table 2), the curves may then be comp unded to give the final curve $\left(\text{intensity}\Big/\frac{\delta\nu}{F_n}\right)$ for the complete line. Scaled plots for H_γ components are shown in figure 16.

In order to determine ion concentrations from equation (4), the curve of observed broadening (intensity/wave number) is fitted on both axes to the calculated curves derived by the calculations described above. Figures 17–20 show typical results using data derived from the photographic work. From a knowledge of the scaling factor on the wave-number axis, it is then easy to find a value of F_n in order that the $\delta\nu/F_n$ axis of the calculated curve should fit the $\delta\nu$ axis of the experimental curve, since F_n is that scaling factor. N is then calculated from equation (4).

For figures 17, 18, 19 and 20 respectively, one unit on wave-number axis corresponds to 12·4, 26·5, 65 and 26·5 cm.$^{-1}$.

The sensitivity of the method may be judged from figure 20, which shows the fitting obtained for $F_n = 120$ kv./cm. in a case where the best fit is obtained for 130 kv./cm. It thus appears possible by this method to measure F_n to $\pm 10\%$. The close agreement between experimental and calculated profiles indicates that absorption effects in the channel are negligible. This was confirmed indirectly by measurements on Zn and Cd vapour, in similar sparks, using the triplet terms in the spectra, whose relative intensities are independent of the mode of excitation. It is hoped to publish these results in a later paper.

Neither the Holtsmark theory, nor that of Hulbert cited above, applies to the undisturbed radiation at the centre of the profile, where the experimental and calculated intensities are widely different, due to the central undisplaced Stark components and the fact that $F \rightarrow 0$ for some of the radiating atoms. The form of the line near the centre of the pattern is probably governed partly by other effects, e.g. a Doppler broadening.

§ 6. ION CONCENTRATIONS IN THE HYDROGEN SPARK CHANNELS EXAMINED

Some results of the photographic technique are summarized in table 3. The peak currents for the 1, 4 and 10 microsec. pulses were respectively 120, 120 and 30 amp.

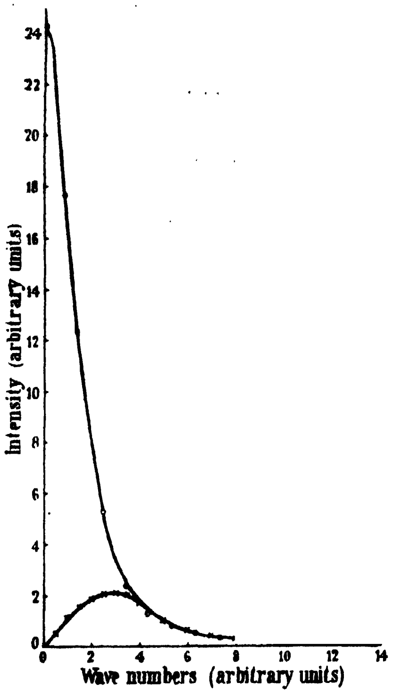


Figure 17. Observed (⊙) and calculated (×) profiles for Hα, 1-microsec. pulse, photographic technique. $F_n = 124$ kv./cm.

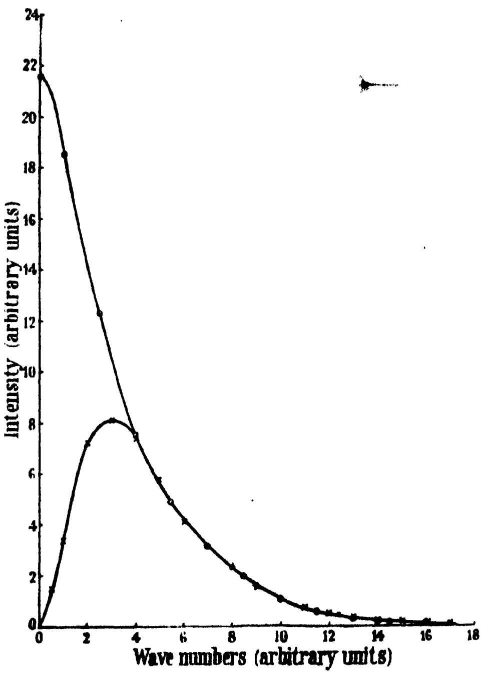


Figure 18. Observed (⊙) and calculated (×) profiles for Hβ, 1-microsec. pulse, photographic technique. $F_n = 133$ kv./cm.

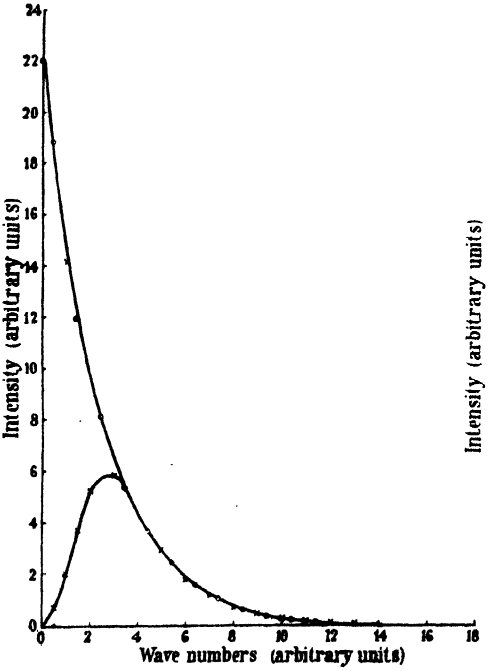


Figure 19. Observed (⊙) and calculated (×) profiles for Hγ 1-microsec. pulse, photographic technique. $F_n = 130$ kv./cm.

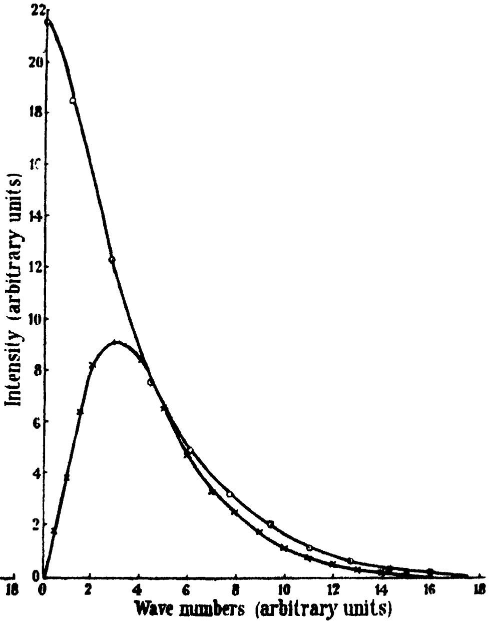


Figure 20. Observed (⊙) and calculated (×) profiles for Hβ, 1-microsec. pulse, photographic technique. $F_n = 120$ kv./cm. Observed results correspond to $F_n = 130$ kv./cm.

Table 3

Plate No.	Pulse length (microsec.)	F_n (kv./cm.)			Mean F_n	N (ions/c.c.)
		Hα	H$_\beta$	H$_\gamma$		
1.5.4	1	124	133	130	130	$2{\cdot}1\times10^{17}$
30.4.3	4	as with 1-microsec. pulse to within about ±5%			130	$2{\cdot}1\times10^{17}$
1.0.3	10	85	79	76	80	$1{\cdot}0\times10^{17}$

Other plates agreed with that given above to within the approximate limits of observation of ±5%over a period of about 9 months, after the technique had been developed and tested, during which time the apparatus was dismantled and reassembled several times for other experiments. Some 40 microphotometer tracings were taken and examined.

The electron multiplier results are exemplified by the data of table 4. 10-microsec. pulses were not used in this part of the work since the spark light was relatively feeble and the narrower lines rendered necessary the use of excessively fine slits.

Table 4

Pulse length (microsec.)	F_n (kv./cm.)		Mean F_n	N (ions/c.c.)
	H$_\beta$	H$_\gamma$		
1	165	163	164	$2{\cdot}9\times10^{17}$
4	as with 1-microsec. pulse to within about ±5%		164	$2{\cdot}9\times10^{17}$

After the technique had been developed, the results were extremely consistent, and records taken over a period of about 4 months agreed within about ±10%.

Since the Holtsmark analysis applies only to the skirts of the line profiles, the photographic measurements in which the total light from the discharges is examined still tend to show the maximum ion concentrations which will be found in the early stages of the passage of current. The electron multiplier results could be taken for any part of the light/time diagram (see figure 22) although the results analysed in detail here refer to the instant of maximum light emission, since the usual method of observation was to measure the peak height on the light/time diagram on the oscillograph screen as the spectrometer wave-length drum was rotated. It was checked, however, that the shape of the light/time diagrams did not change as the line under examination was traversed across the multiplier slit; hence, for example, the extreme skirts of the Stark profile (corresponding to the highest fields encountered during the whole "visible" discharge) did not occur especially at the beginning of light emission.

Since, also, the photograph and multiplier plots for 1 and 4 microseconds agree extremely closely in shape, it seems likely that the field, or ion concentration, changed very little during the observed part of the discharge. This subject will be discussed below in detail. The conclusions from the above argument appear to be that (*a*) the ion concentration varies little during the time of discharge in which observations are made, and therefore (*b*) the ion concentrations measured by the two methods should agree. Hence from (*b*), the difference in values of *N* for the same discharges, see tables 3 and 4, is due to some unexplained consistent error in one of the measurements, since both methods give accurately repeatable results ($\sim \pm 10\%$ error) after all precautions have been taken. The error in *N*, which is $2{\cdot}5 \pm 0{\cdot}4 \times 10^{17}$ ions/c c., is, however, small for this kind of work, and the results appear to be the most accurate so far published. Since the electron multiplier gave the same results for 1- and 4-microsec. pulses of the same peak current, as would be expected, it seems unlikely that its frequency response was inadequate. The circuits were carefully tested for response (§ 2). It is considered more likely that the photographic results are in error since other work (to be published separately) indicated that the spectral response of the plates, particularly as regards intermittency effects (see, for example, Mees, 1944), might not be as consistent and predictable as existing theory indicates. There appear to be little published data on the response of plates to light pulses lasting 1–5 microsec. The discussion in §§ 2 and 3 shows that all the usual precautions were taken in the photographic work (Sawyer, 1944).

The implications of conclusion (*a*) are of interest. It is immediately clear (Lawrence and Dunnington, 1930, and others) that the ion concentration probably varies greatly during the early parts ($\sim 10^{-7}$ sec.) of spark discharges. Lawrence and Dunnington obtained $\sim 30\%$ ionization of Zn vapour in $\sim 10^{-4}$ sec. from the commencement of discharge, corresponding to a temperature of $\sim 13{,}000^\circ$. It seems probable that Saha's equation will not apply to 1–10 microsec. sparks, in which case spark temperatures deduced in that way from measured ion concentrations will be inaccurate. In the present case ($T \sim 12{,}000^\circ$ from $N = 2{\cdot}5 \times 10^{17}$ ions/c.c.), Craggs and Meek (1946), figure 14(1), for example, shows that the light emission from the hydrogen sparks is negligible for perhaps the first 0·25 microsec. of discharge. The current reaches its maximum value in $\sim 0{\cdot}1$ microsec., so that the discharge is passing the full current before visible light emission is appreciable. The present experiments are not intended to refer to this early period of the discharge, which will be treated in later work. It is known, however, that the voltage gradient in the present type of spark changes greatly only during the first 0·5 microsec. of conduction (Flowers, 1943; Craggs, Haine and Meek, 1946; and others) and that the ratio of the numbers of ionizing and exciting collisions will vary greatly, for that reason, during the early part of the spark. Thus Penning (1938), using infinitely small currents, showed that between X/p 20 v./cm./mm. and $X/p \to 0$ (which are the approximate values for a hydrogen spark at the beginning and end of the high field period) the ratio of electron energy spent in exciting and ionizing was rapidly increasing, i.e. more light was being emitted at low X/p.

It is perhaps fortunate that the early part of the discharge is not recorded with visible light, since that fact enables some information to be deduced for the afterglow period, i.e. the time, after the current has fallen to zero, within which light is still emitted. Craggs and Meek (1946) discussed mechanisms of afterglows in hydrogen and argon, and it was suggested (*loc. cit.*) that the light could only be due to recombination if $\alpha \sim 10^{-11}$ (α is the recombination coefficient for electron ion collisions). Whilst recombination certainly takes place in the hydrogen sparks, as shown by the series limit continuum of figure 4, it appears from the present experiments, which show that the ion concentration varies little during the whole discharge (since the multiplier and photographic Stark profiles are identical in shape, see figures 9 and 10), that electron/ion recombination must indeed be slight, and hence that $\alpha < 10^{-11}$ for the spark channels. It is likely that the ion concentration during the afterglow has dropped by <20%, i.e. $\alpha \sim 2 \times 10^{-12}$ (see table 5 and discussion, both in Craggs and Meek (1946)).

Massey (1938) and Bates *et al.* (1939) quote values of the total cross-section for radiative electron/ion recombination,

$$Q_e = \Sigma_n Q_e^n, \qquad \ldots\ldots(8)$$

where Q_e^n is the coefficient for recombination into the nth state; $\sum_1^\infty Q^n = 23 \times 10^{-21}$ cm² for an electronic energy V of 0·28 volt. The cross-section varies approximately as 1/2. Assuming an electron temperature in the spark channel of 15,000°, then $V \sim 2$ volts, so $v \sim 8 \times 10^7$ cm./sec. Since

$$\alpha \doteq vQ_e, \qquad \ldots\ldots(9)$$

then $\alpha \sim 8 \times 10^7 \times 3 \times 10^{-21} \sim 2 \times 10^{-13}$. To this must be added $Q_e^0, \sim 6 \times 10^{-22}$ cm² for 2-volt electrons (Massey 1938, p. 35) and the total value of α is $\sim 10^{-13}$ for the spark channel. Radiationless three-body collisions (Smith, 1936) can be shown to have a cross-section about equal to that for radiative capture if the electron concentration is $\sim 10^{18}$/c.c. Smith suggests that such concentrations are unlikely to be encountered in electrical discharge experiments, but an exception could apparently be made for spark channels of higher current than those described here. It is hoped to re-examine recombination problems in spark channels, particularly for the highest attainable ion concentrations.

In the earlier paper (Craggs and Meek, 1946), the ion concentration for the present sparks (120-amp. 1–4-microsec. pulses) in hydrogen was deduced by several methods as being approximately 10^{17} ions/c.c., in excellent agreement with the present results. In particular, attention is here directed to the method of equations (2) and (3) of the early paper, in which the ion concentrations is found from a knowledge of channel radius (taken as 0·75 mm.) and voltage drop. The latter, about 90 v./cm., corresponded to $N \simeq 6 \times 10^{17}$ ions/c.c.

In order to obtain preliminary information on channel structure, some photographs of single 1-, 2- and 4-microsec. sparks were taken and the channel images traced on the micro-photometer. Results for 1- and 4-microsec. sparks are shown in figures 21 and 22. The full diameters are about 1·3 mm. for both cases, which would give $N \simeq 4 \times 10^{17}$ ions/c.c. from the other data published (Craggs and Meek, 1946). The fact that the ion concentration is much less for the lower current 10-microsec. discharges (table 3) is of interest. Since the 1- and 4-microsec. sparks give the same values of N, it seems unlikely that the length of the 10-microsec. current pulse, which might allow expansion of the spark channel, is the controlling factor. It is suggested that the spark channel expands less rapidly than would be necessary to give constant-current density (or ion concentration), so that the higher current discharges are relatively more constricted (C. J. Flowers, 1943). However, figures 21 and 22 show that the channels are non-uniform in luminosity and are brighter at the centre. Although the relation between intensity of light emission and ion concentration is not yet accurately known, it is reasonable to suggest a correspondingly high radial ion-concentration, in which case N, as determined from voltage drop, would need further correction. Figures 21 and 22, which were taken in order to show that the spark channels expanded inappreciably between 1 and 4 microsec. after their time of origin, could owe their shapes to the mechanism of growth of a channel from a streamer, in the sense that the photographic record integrates an infinite number of succeeding stages of the spark from its beginning as a streamer thin compared with the spark channel which succeeds it (we are particularly indebted to Professor J. M. Meek for discussions on this topic), in which case the postulation of radial charges in ion concentration would require verification. It is considered, since the growth of spark luminosity is so slow (see figure 22) with visible light, that this alternative explanation is not likely, and that figures 21 and 22 indicate strong radial changes in N at the times 1-4 microsecs. It is intended that the technique illustrated by figures 21 and 22 shall be developed for other spark discharges.

§ 7. CONCLUSIONS

The ion concentration in certain hydrogen spark channels has been measured and found to be about $2{\cdot}5 \times 10^{17}$ ions/c.c. by observations of the Stark broadening of the Balmer lines H_α, H_β and H_γ. The analysis takes full account of the fine structure of the above lines. It is deduced from observations of the Stark profiles with an electron multiplier technique that the electron/ion re-combination coefficient in the experimental conditions is $\sim 2 \times 10^{-12}$. The spark temperature, as deduced from Saha's equation, is about 12,000°.

§ 8. ACKNOWLEDGMENTS

The authors wish to express their thanks to Mr. F. R. Perry, in whose laboratory the work was carried out, for his continuous interest and support. The authors are also indebted to Professor J. M. Meek (University of Liverpool) for many stimulating discussions, to Dr. S. Tolansky (University of Manchester) for the use of a microphotometer, and to their colleagues Mr. G. J. Scoles and Mr. C. J. Braudo for advice and help with circuit problems. Thanks are due to Sir Arthur P. M. Fleming, C.B.E., D.Eng., Director of Research and Education, and Mr. B. G. Churcher, M.Sc., M.I.E.E., Manager of Research Dept., Metropolitan-Vickers Electrical Co. Ltd., for permission to publish this paper.

REFERENCES

Bates, D. R., Buckingham, R. A., Massey, H. S. W. and Unwin, J. J., 1939. *Proc. Roy. Soc.*, A, **170**, 322.

Brainerd, J. G., Koehler, G., Reich, H. J. and Woodruff, L. F., 1942. *Ultra High Frequency Techniques* (London: Chapman and Hall).

Craggs, J. D., Haine, M. E. and Meek, J. M., 1946. *J. Inst. Elect. Engrs.*, **93**, IIIA, 936.

Craggs, J. D. and Meek, J. M., 1946. *Proc. Roy. Soc.*, A, **186**, 241.

Finkelnburg, W., 1931. *Z. Phys.*, **70**, 375.

Flowers, J. W., 1943. *Phys. Rev.*, **64**, 225.

Forsythe, W. E. and Adams, E. Q., 1945. *J. Opt. Soc. Amer.*, **35**, 108.

Fucks, W. and Bongartz, H., 1943. *Z. Phys.*, **120**, 468.

Gebauer, R. and von Traubenberg, H. R., 1930. *Z. Phys.*, **62**, 289.
de Groot, W., 1931. *Physica*, **11**, 307.
Holtsmark, J., 1919. *Ann. Phys., Lpz.*, **58**, 577.
Hulburt, E. O., 1923. *Phys. Rev.*, **22**, 24.
Hulburt, E. O., 1924 a. *Phys. Rev.*, **23**, 106.
Hulburt, E. O., 1924 b. *Astrophys. J.*, **59**, 177.
Hulburt, E. O., 1926. *J. Franklin Inst.*, **201**, 777.
Lawrence, E. O. and Dunnington, P. G., 1930. *Phys. Rev.*, **35**, 396.
Loeb, L. B. and Meek, J. M., 1940. *J. Appl. Phys.*, **11**, 438.
Margenau, H. and Watson, W. W. 1936. *Rev. Mod. Phys.*, **8**, 22.
Massey, H. S. W., 1938. *Negative Ions* (Cambridge University Press).
Mees, C. E. K., 1944. *The Theory of the Photographic Process* (New York : Macmillan).
Merton, T. R., 1915. *Proc. Roy. Soc.*, A, **92**, 322.
Minkowski, R., 1929. *Handbuch der Experimental Physik*, **21**, p. 389 *et seq.* (Leipzig : Springer).
Pannekoek, A., 1937. *Mon. Not. R. Astr. Soc.*, **93**, 694.
Penning, F. M., 1938. *Physica*, **5**, 286.
Raether, H., 1939. *Z. Phys.*, **112**, 464.
Sawyer, R. A., 1944. *Experimental Spectroscopy* (New York : Prentice-Hall).
Smith, R. A., 1936. *Proc. Camb. Phil. Soc.*, **32**, 482.
Verweij, S., 1936. *Publ. Astr. Inst. Univ., Amsterdam*, No. 5,
Weisskopf, V., 1933. *Phys. Z.*, **34**, 1.
White, H. E., 1934. *Introduction to Atomic Spectra* (New York : McGraw-Hill), p. 402.

ELECTRON/ION RECOMBINATION IN HYDROGEN SPARK DISCHARGES

By J. D. CRAGGS and W. HOPWOOD,
Metropolitan-Vickers Electrical Co. Ltd., Manchester

MS. received 10 *Janaury* 1947

ABSTRACT. The afterglows following short sparks in hydrogen at a pressure of one atmosphere have been described in earlier publications. The present communication presents later results and a detailed analysis of the afterglows. It is shown that an electron/ion recombination process is apparently the cause of the light emitted in the afterglow. On this basis the appropriate coefficient of recombination has been calculated. The implications of the results, and relevant literature, are discussed.

§ 1. INTRODUCTION

Experimental data relating to electron/ion recombination are very scarce, and some interest is attached to them. Further efforts to obtain recombination cross-sections are also to be encouraged, since the results would be of considerable fundamental and technical interest. For example, the de-ionization of gases following electrical discharges in them depends partly on the recombination processes and in certain cases (for example, in high-temperature discharges in hydrogen) the interacting particles are electrons and positive ions, since negative ions and molecules can often be shown to be relatively insignificant.

Recombination data are used also in studies of the upper atmosphere, and their importance in that field is stressed in the Gassiot Committee's report (1942/43).

Attention has recently been directed to the existence of afterglows following spark discharges, in high pressure gases (Meek and Craggs, 1943; Rayleigh, 1944; Craggs and Meek, 1945 and 1946) and the possibility of determining electron/ion recombination coefficients from such observations has been stressed by Craggs and Meek (1946). The results thus obtained refer to the special conditions obtaining in such discharges, and their relevance to other problems can only be decided after a study of such conditions.

§ 2. DISCUSSION OF PREVIOUS WORK

Many excellent investigations of recombination between positive and negative ions have been made (Sayers, 1938, and references there cited), but experimental data on electron/positive ion recombination are scarce, as has been pointed out in § 1.

Mohler (1937) found that the recombination coefficient α, defined (where t is time and N is the number of ions or electrons, assumed equal, per cc. of gas) by the equation

$$dN/dt = -\alpha N^2, \qquad \ldots\ldots(1)$$

could be determined for an interrupted Cs vapour glow-discharge as $3{\cdot}4 \times 10^{-10}$. The discharge pressure was $10-33 \times 10^{-3}$ mm. Hg, and the electron temperature about 1200°. This paper followed earlier reports by Mohler (1928 etc.) on similar work; attention is particularly directed to the experiments of Mohler and Boeckner (1929) on the recombination spectra of ions and electrons in Cs and He. The great increase in the recombination coefficient with decreasing electron energy is there emphasized. The above data, and those of Kenty (1928) for the afterglow of an argon arc, may be summarized by stating that $\alpha \sim 10^{-10}$ for the conditions studied.

There seem to be no published results for hydrogen, but Mohler (1937) has discussed the case using the theory of Stueckelberg and Morse (1930). It appears that when the kinetic energy of the electrons is small relative to that of the level into which the electron falls, the recombination coefficient α_n for that level n is given by

$$\alpha_n = 5{\cdot}94 \times 10^{-13} A_n V^{-1/2}, \qquad \ldots\ldots(2)$$

where A_n is a simple function of n for $n \geqslant 3$. V is the electron energy in electron-volts. When the electron's kinetic energy is much greater than that of the level n, Oppenheimer (1928) gives

$$\alpha_n = 6{\cdot}3 \times 10^{-11}/V^2 n^3. \qquad \ldots\ldots(3)$$

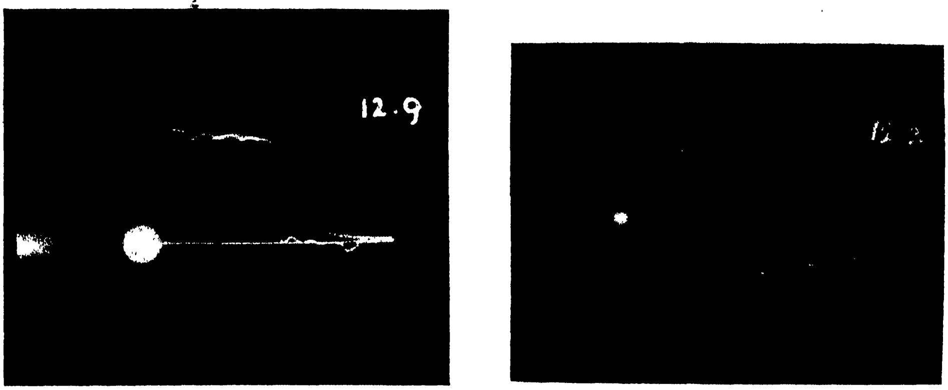
Equation (2) is used for $n < n_0$, where

$$n_0 = (13{\cdot}54/V)^{1/2},$$

and equation (3) is valid for $n \gg n_0$ although it may be used when $n = n_0$ to give approximate results.

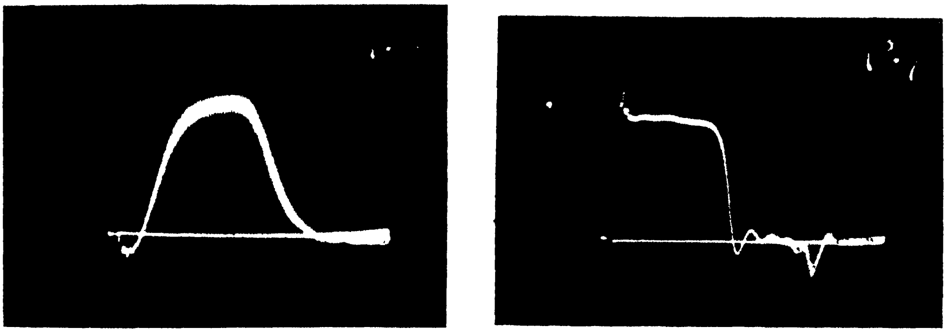
Take as an example $V = 0{\cdot}1$ ev. Then $11 < n_0 < 12$, and total α (i.e. $\Sigma\alpha_n$) is $2{\cdot}3 \times 10^{-11}$, but the errors involved are probably large and are not assessed in detail by Mohler (*loc. cit.*). The values of A_n published by Stueckelberg and Morse (1930) were used for these calculations.

More recently, the observations by Lord Rayleigh on long duration Balmer series emission in hydrogen (1944) led Zanstra (1946) to calculate the electron/ion



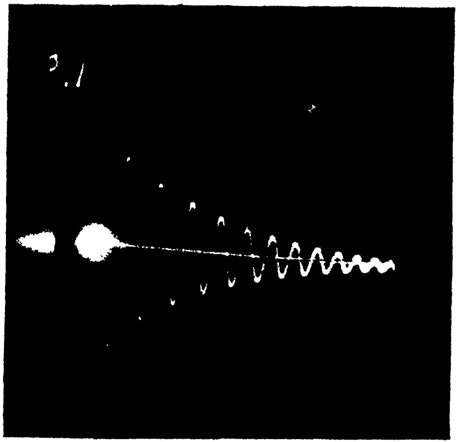
(a)
4-microsec. 120-amp. pulse. (Light/time and current oscillograms.)

(b)
Time base for (a).



(c)
Light/time oscillogram for (d).

(d)
10-microsec. 30-amp. pulse-current oscillogram.



(e)
Time base for (c) and (d).

Figure 1. Oscillograms of current, light emission and time-base speeds. The timing oscillation has a period of 2·8 microsec.

recombination cross-sections for Rayleigh's experiments, making certain reasonable assumptions in the process. The results will be discussed more fully in § 4, but an electron temperature of 1000°, corresponding to an average electron energy (Maxwell distribution) of about 0·13 ev., gives $\alpha = 1{\cdot}2 \times 10^{-12}$.

§ 3. EXPERIMENTAL TECHNIQUES

Mohler and Kenty (references given in § 1) were able to use mechanical means for stopping the flow of current in their discharges which were arcs or glows of long duration, i.e. by the use of commutators short circuiting the terminals of the discharge tube. The decay of ionization was then studied in the absence of current by various means such as probes.

This technique is clearly impracticable for spark discharges lasting only a few microseconds, but the utilization of square current pulses (~100 amp. peak), following normal radiolocation practice, leads to the possibility of observing the short hydrogen afterglows and deriving approximate values for recombination coefficients from them. The circuits have been briefly mentioned by Craggs and Meek (1946) and will be described in more detail in a forthcoming publication by Craggs, Haine and Meek (1946). A typical current pulse (10 microsec., 30 amp.) is shown in figure 1 (*d*) and a typical light emission/time record, figure 1 (*c*), indicates that a short afterglow is observable. Figure 1 (*c*) was taken with a photoelectric electron multiplier tube and cathode ray oscillograph.

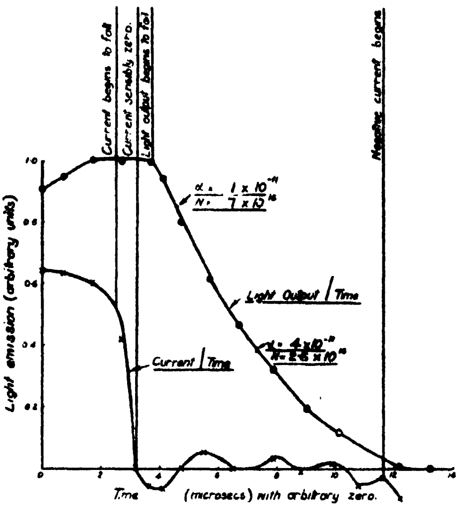


Figure 2. Analysis of hydrogen afterglow (*a*) 10 microsec. 30 amp. pulse.

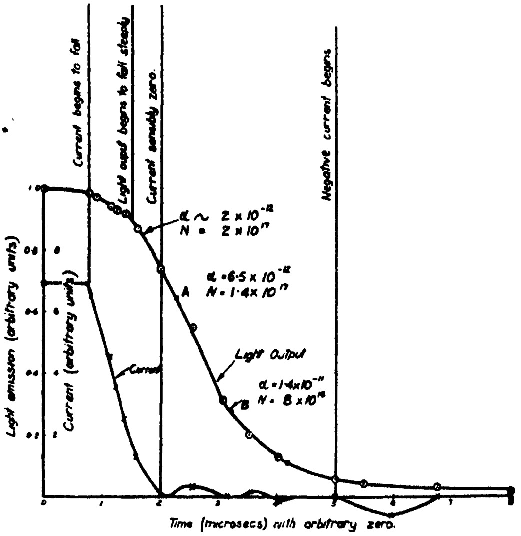


Figure 3. Analysis of hydrogen afterglow (*b*) 4 microsec. 120 amp. pulse.

The current pulse of figure 1 (*d*) (reproduced in figure 2 after a correction by calibrations of the photo-tube and oscillograph), which is produced with an artificial transmission line arranged in a suitable discharging circuit, shows a fall-off of current taking a finite time, i.e. about 0·5 microsec. This current therefore vitiates measurements of the afterglows, but only for perhaps the first 20% or less of its life (figure 3) if a suitably matched transmission line is used for pulse generation.

It will be noticed from figure 1 (*a*), (*c*) and (*d*) that the peak light-output (visible radiation) is reached some time after the current has begun to fall. This is found by careful fitting of the current and light output oscillograms at the beginning of the current pulse, and hence appears to be a real effect. A possible explanation is that, during current flow, the electrical energy is stored in the discharge in the form of positive ions, fast electrons and possibly highly excited atoms, and that this distribution needs a finite time ($\sim$1 microsec.) to establish some form of transient equilibrium at a lower electron temperature. Some support for this argument may be derived qualitatively from, e.g., figure 1 (*a*), which shows that, even when full current is established, the maximum light-emission is not reached for some further 3 microsec. or more.

Penning (1938) and others have studied similar effects for low-pressure glow discharges for vanishingly small currents, where the problem is much simplified by the time stability of the discharge. However, it is hoped to obtain information from further experiments on short-time sparks bearing on this interesting problem of the variation of electron temperature with time.

It is important to consider the frequency response of the amplifier/photo-tube arrangement. The high-frequency amplifier has been briefly described by Craggs and Hopwood (1946). The circuit time-constant was about 0·3 microsec., i.e. was low compared with an afterglow time of 3–4 microsecs. (15,000-Ω output load, and 20 mmF. oscillograph input capacitance) and the light/time diagrams were not sensibly altered in shape when the load was decreased to 3000 Ω, but the latter load gave inconveniently low output voltages. The frequency response of the amplifiers was checked with pulse generators and figures 4 (*a*) and (*b*) show typical results.

The response of the multiplier tube to light pulses of known shape is much more difficult to determine, and would be limited by transit-time effects, or a lag in the photo-emission mechanism. Zworykin *et al.* (1936) show these to be negligible for ordinary photocells at, e.g., 1 mc./sec., which may be considered as an effective frequency for the present work. R. A. Houstoun (1936) showed that with a standard commercial vacuum photocell the difference in response for exposures of $1{\cdot}59\times10^{-6}$ and $7{\cdot}4\times10^{-8}$ sec. was only about 0·1%. Further experiments cited by Houstoun showed that the response from $4{\cdot}4\times10^{-3}$ sec. and $1{\cdot}47\times10^{-7}$ sec. exposures were the same within the limits of experimental error. Recent work by Geist (1941) proved that even for 10 cm. diameter photocells at only 150 v., the transit time effect was negligible for operating frequencies of 1 mc./sec.

The authors performed some experiments with a rotating mirror and slit system arranged to give short light pulses, but a more satisfactory and simpler way exists of showing that the multiplier tubes are capable of responding to light changes occurring more rapidly than those in hydrogen afterglows, viz. the rate of increase of light output from argon sparks, during the initial stages of the spark, is seen to be considerably greater than that for hydrogen sparks when examined with electron multipliers (Craggs and Meek, 1946).

§ 4. EXPERIMENTAL RESULTS

Many new data have been collected since the earlier work was reported (Craggs and Meek, 1946). Typical records are shown in figures 1 and 2, and afterglow curves taken with two different sets of apparatus more than two years apart are shown superimposed in figure 5. Figure 5 (*a*) shows a record from the previous research (Craggs and Meek, *loc. cit.* in which no amplification between multiplier and oscillograph was used) compared with one of the new oscillograms taken with a different type of electron multiplier and with the introduction of an amplifier (figure 5 (*b*)).

Now, Craggs and Hopwood (1946) have shown that the ion concentration for the hydrogen spark channels used in these experiments is 2·5 × 10^{17} ions/cc. (N_{i_0}), i.e. at the commencement of the fall of light emission. The fact that such high ion concentrations (giving a pronounced Stark effect) are found suggests at once that calculations of recombination coefficients will apply only approximately to spark channels, i.e. to an extent depending on the importance of the "level smearing". It is hoped to discuss this matter in more detail elsewhere. Since the recombination coefficient α is given by

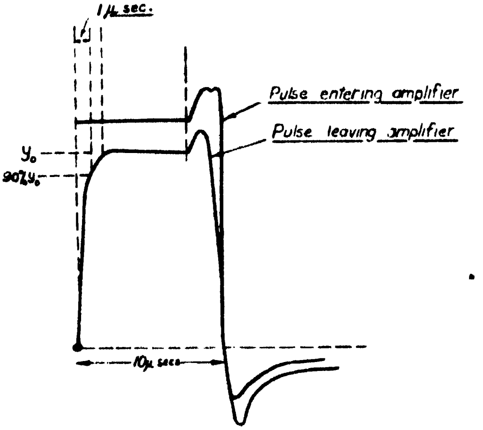


Figure 4. Frequency response of one of the amplifiers used with the photo-tube and oscillograph system. The amplifier gives ~90% in microsec.

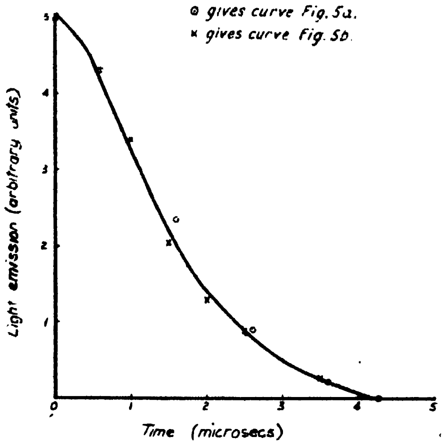


Figure 5. Afterglows as measured with two sets of recording equipment.

$$dN_i = -\alpha N_i^2 dt, \qquad \ldots\ldots(4)$$

which is equation (1) re-written, and if it is assumed that recombination is responsible for the afterglow, then the light output L_t at time t is given by (k is a constant)

$$L_t = k \, . \, \frac{dN_i}{dt} = k \, . \, \alpha N_i^2. \qquad \ldots\ldots(5)$$

The positive ion and electron concentrations are supposed equal, since diffusion of electrons must be negligible (see below). Thus

$$\frac{1}{N_{i_1}} - \frac{1}{N_{i_0}} = \alpha \, . \, \Delta t,$$

where Δt is the time interval during which N_{i_0} (the ion concentration with current flowing) falls to N_{i_1}. Assuming α constant during Δt, which is taken as 0·5 microsec., then since $L_t = KN_i^2$, N_{i_1} is found from N_{i_0} and α is determined. A more accurate method of determining α at various points on the afterglow curve is by graphical integration, using the following procedure. When the light has fallen sensibly to zero, as it does for example at $t=13$ microsec. in figure 2, then the ion concentration at that time may be taken as zero. The area of the curve included from an earlier time, when the ion concentration is say N_{i_1}, to the time when the ion concentration is zero, is (see equation (5))

$$k\int_{t_{N_i}=0}^{t_N} \frac{dN_i}{dt} = kN_i.$$

This integral taken from the beginning to the end of the afterglow is kN_{i_0}, where N_{i_0}, is known, and so k may be found. Thus N_i can be found at any point in the afterglow. Finally, since the ordinate at any point is $k\alpha N_i^2$ with the values of α and N_i obtaining at that point, α may be found since it is then the only unknown. It was in this way that the values of α shown in figures 2 and 3 were determined.

Figures 2 and 3 show afterglow curves in which α has been calculated for the marked points, and α is shown to rise (figure 3) from about 2×10^{-12} to 1×10^{-11} in about 2 microsec. As mentioned above, the persistence of a small current for the first period (perhaps one-third) of the afterglow would vitiate the results there. The records of figures 1, 2 and 3 are, however, not affected in this way. These values of α are total recombination cross-sections if the fraction of recombinations giving Balmer quanta is approximately independent of temperature over the range, say, 1000–5000° c. This assumption seems reasonable (Zanstra, 1946; Cillié, 1932, 1946).

§ 5. DISCUSSION OF THE EXPERIMENTAL RESULTS

It is next necessary to justify the assumption that electron/ion recombination is the controlling process for the hydrogen afterglows. Molecular hydrogen is apparently absent from the discharge (100% dissociation) because no molecular spectrum is observed; the electrical and thermal properties of the spark channels are such (Craggs and Meek, 1946) that this result, in accordance with Lord Rayleigh's experiments (1944), is to be expected. It is not certain that negative ions are absent in the hydrogen sparks but this difficulty is eliminated by showing that the afterglow curves of figures 1 etc. are those which are, in fact, given by the continuous radiation on the short wave-length side of the Balmer series limit at 3647 A., as well as by the visible radiation in the Balmer series, since the series limit continuum is due to electron/ion recombination.

Loeb (1939, pp. 144, 145 and elsewhere) has discussed the other forms of recombination, i.e. preferential, columnar and initial recombination. It is unlikely that these processes are relevant to the case of spark channels in hydrogen, and so volume electronic recombination is the only mechanism of importance.

A Wratten filter No. 1 passes radiation for $\lambda > 3600$ A. (95% transmission at $\lambda=3800$ A. and 0·1% transmission at $\lambda=3650$ A.) and the spark radiation for $\lambda < 3650$ A. can thus be found by measuring the total radiation and that passing through the filter and performing a subtraction. Some results are shown in figures 6 and 7, and it is clear from the latter that the recombination radiation decays at the same rate as the visible radiation in the Balmer series. The subtractive process is rarely used since it is more convenient to observe the visible radiation. Later records taken with a spectrometer and sensitive amplifier show that H_α, H_β and H_γ all decay at sensibly the same rate, which is that of the Balmer continuum in figure 7.

In order to calculate the recombination coefficient it is not necessary to show that all the afterglow radiation is due to recombination, but it would clearly be of interest to do so, since the relative importance of other processes (e.g. thermal excitation) could then be assessed.

It is fortunate that astrophysicists have studied the hydrogen recombination spectrum in great detail, and it is interesting to apply the results of their calculations to the present case. Pioneer work in this field is due to Zanstra (1927). Menzel and Cillié (1937) show that E_c, the energy emitted (frequency ν) per second per c.c. of gas for a unit wave number range in the continuum, is given by

$$E_c \propto e\,\frac{\chi_2 - h\nu}{kT_e}. \qquad \ldots\ldots(6)$$

χ_2 is the ionization potential from the second quantum state, i.e. approximately 3·38 v., k is Boltzmann's constant and T_e is the absolute electron temperature.

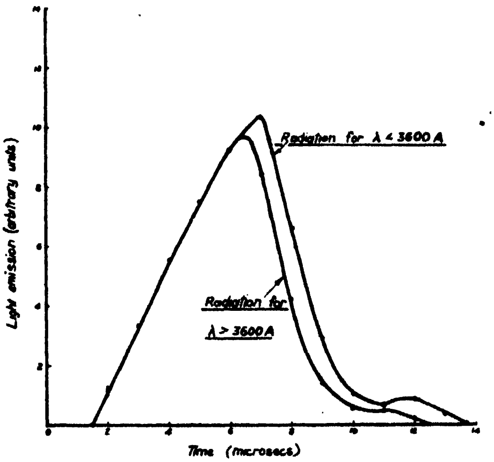


Figure 6 (*a*). Radiations emitted from a hydrogen spark, above and below the Balmer series limit. The curves have scaled together at the 6 microsec. point.

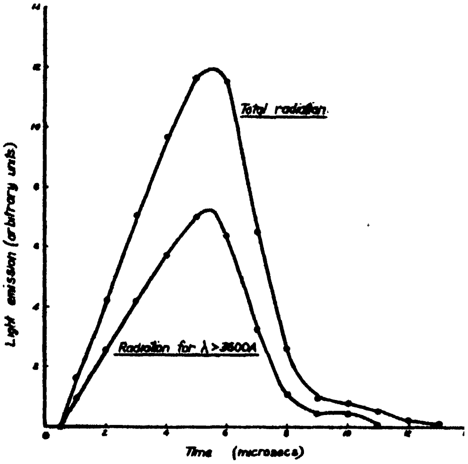


Figure 6 (*b*). The light emission from a hydrogen spark, taken with and without a Wratten No. 1 filter between source and photo-tube.

For $T = 20000°$, equation (3) gives the results plotted in figure 8, where it is seen that E_c varies only by about 10% from the mean at about 3500 A., over the range from the series limit at 3647 A. down to 3300 A. which is about the cut-off limit of the glass optical system used in the present experiments. This range covers some $2{\cdot}5 \times 10^3\,\mathrm{cm}^{-1}$. Cillié (1936) calculated the strength of H_β (and the other Balmer lines for a recombination spectrum) in terms of the energy emitted per unit wave number at the series limit and found this ratio to be about 10^3 at $T_e = 20000°$. The intensities of 8 Balmer lines (excluding H_α, which is much attenuated by the extremely low multiplier response for red light) are plotted in figure 9. It is easily shown that the total continuous radiation is probably ~50% of the total Balmer radiation, making allowance for the colour response of the multiplier. The latter is discussed in a forthcoming paper by the present authors.

Figures 7 and 8 show that the continuous radiation is of this order when allowance for the multiplier's colour-response is made. This approximate experiment which is to be extended in a separate investigation, strongly suggests that the hydrogen afterglows are due almost entirely to recombination radiation.

Zanstra (1946) has proposed, as Craggs and Meek (1945 and 1946) had done, that Rayleigh's observations of long duration Balmer spectra should be explained by recombination, and worked out the electron/ion recombination coefficients from the theoretical results of Cillié (1932) and others, giving the data of table 1.

Table 1

T (° c.)	1000	5000	10 000	20 000	50 000
$10^{14}\,C$	218	74	46·0	27·0	13·5
$10^{14}\,C'$	122	33	17·7	9·4	3·6

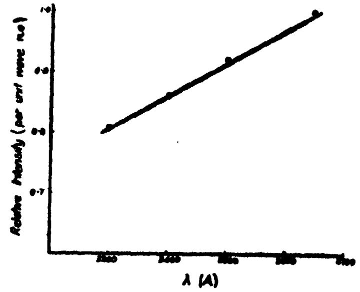


Figure 7: The intensity distribution in the Balmer series limit continuum for $T_e = 20\,000°$.

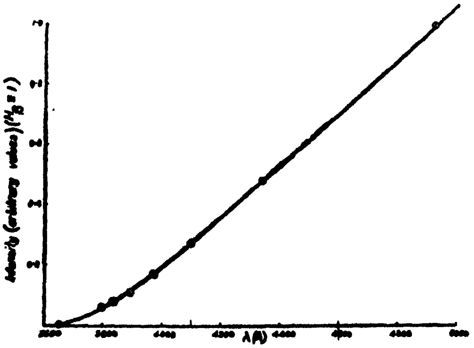


Figure 8. Intensity distribution in Balmer series (recombination) from Cillié's data. $T_e = 2000°$.

C is given by

$$dN/dt = -N^2 C, \qquad \ldots\ldots(7)$$

where

$$C = \sum_{n=1}^{n=\infty} C_n, \qquad \ldots\ldots(8)$$

and n refers to a particular energy level,

$$C' = \sum_{n=3}^{n=\infty} C_n, \qquad \ldots\ldots(9)$$

and gives (with equation (7)) the number of recombinations/c.c./sec. giving quanta emitted as Balmer radiation.

Comparison of our results (figure 3) with Zanstra's computations show that the electron temperature probably falls from its value of 10 000° in the spark to $<1000°$ in less than a microsecond. The rough estimate of spark temperature just given is based on various experiments (Craggs and Meek, 1946, and references there cited). For example, experiments in this laboratory (Craggs, Haine and Meek, 1946) have shown that for 120 amp., 4 microsec. argon sparks at about 1 atmos. pressure X/p falls to about 0·1 v./cm./mm. Hg at the end of the period of current flow.* If the data of Allen (1937) can be considered relevant, which is probably only approximately true, T_e is then 10 000°, with higher values up to

---

* X is the electric field.

$>100\,000°$ in the preceding parts of the discharge. The fall may take place during current flow, since (Craggs, Haine and Meek, 1946) X/p falls from an initial value of ~ 40 v./cm./mm. to <1 v./cm./mm. whilst current is still flowing in a spark in which a 4 microsec. square current pulse is used, or it may have taken place after the current has fallen to a very low value.

It is of great interest to note that from figures 2 and 3 the recombination coefficients for comparable periods* in the afterglow are nearly the same despite the great difference in the initial values of ion concentration (the latter quantities were measured in a separate research and details will shortly be published). The conclusion from this finding is that the electron temperature is largely dependent on ion concentration for sparks of the type described here. Further experiments on this and other aspects of spark channels are in progress. It is interesting also to note (figure 2) the slower afterglow decay for the lower value of initial ion concentrations, in accordance with expectations. The diffusion of electrons out of the spark channel in the times involved here can be shown by approximate methods to be negligible (Margenau *et al.*, 1946, and Loeb, 1939, p. 175). Because of the large number of positive ions present, diffusion will be at a slower rate than that for free electrons and the movement should certainly be $<0{\cdot}1$ mm. in 4 microsec.

Zanstra points out that since the cross-section for electron excitation is $\sim 10^{-16}$ cm.^{2} and that for capture is $\sim 10^{-21}$ cm.^{2} the former process will predominate in the earlier part of the de-ionization process. By virtue of this, however, the electronic kinetic energy is soon spent and the recombination process then assumes control.

An illustration is provided by elementary consideration of energy loss at electronic collisions, for which the tables of Healey and Reed (1941) have provided the numerical data. It must be emphasized that conditions in spark channels are so transitory and ill-defined that the following treatment is necessarily approximate. At $X/p = 0{\cdot}25$ v./cm./mm. the mean electronic velocity is 2×10^7 cm./sec. in hydrogen, with a mean free path of

$$\frac{3{\cdot}6 \times 10^{-2}}{760} \text{ cm.}$$

at $p = 760$ mm. Hg. The spark channel temperature is much higher than room temperature but in the times involved (few microsec.) gas movement is negligible and so the gas density remains constant. In 1 microsec. the number of electronic collisions is therefore

$$\frac{2 \times 10^7 \times 10^{-6} \times 760}{3{\cdot}6 \times 10^{-2}} \sim 4 \times 10^5.$$

The fraction of the energy lost per collision is (Townsend and Bailey, *loc. cit.*) 26×10^{-4}, so the total energy loss in 1 microsec. is thus $\gg 100\%$ or, at least, is very great. It is therefore clear that a rapid fall in electron temperature, such as that indicated by the results shown in figure 2 etc. is reasonable. Experiments intended to determine the electron temperatures in these sparks are in progress.

Since this paper was first written, Margenau *et al.* (1946) have studied the dissipation of energy of an electron swarm in T.R. switches (radar practice). They show by similar arguments to the above that, in argon at 10 mm. Hg pressure,

---

* i.e. when the ion concentrations are equal.

free electrons will lose most of their energy in ~4 microsec. At 760 mm. Hg pressure, the relaxation time would then be <0·1 microsec.

The importance of recombination processes in determining electrical conditions in the earth's upper atmosphere is well known. Appleton (1937) has summarized the situation and has pointed out that from the data of Morse and Stueckelberg (1930) $\alpha = 1{\cdot}1 \times 10^{-11}$ for $T = 400°$ K. (0·05 ev. energy of electrons) and for capture to the ground state; α varies as T^{-1}. The contribution to α due to recombination into excited states does not appreciably increase α. Appleton showed that α deduced from Milne's work cited by Chapman (1931) is in agreement with that already given, and thus that there certainly appears to be a discrepancy between theoretical and experimental values of α. Sayers (1943) pointed out that α (F_2 region) is apparently $\sim 10^{-10}$ cm.³/sec., in agreement with the existing low-pressure data (Mohler *et al.*), but emphasized that conditions in the two sets of circumstances may not be comparable. Massey and Bates (1943) discuss the electron/ion recombination of oxygen also with reference to conditions in the ionosphere and show that, for example, at 1000° K. the total calculated coefficient is only $\sim 1{\cdot}5 \times 10^{-12}$ cm.³/sec., which is much lower than the experimental values of Mohler and Kenty and those determined for the F_2 layer. Massey and Bates stress the need for further investigations.

Bates *et al.* (1939) have worked out some useful expressions for recombination processes used by Craggs and Hopwood (1946), where it is shown that α (hydrogen) is about 2×10^{-13} cm.³/sec. for an electronic energy of 2 ev. ($T_e \sim 15\,000°$). Using the same expressions, the values of α at 5000° and 1000° respectively are about 5×10^{-13} and 10^{-12}, thus agreeing satisfactorily with Zanstra's data given above.

§ 6. CONCLUSIONS

It would appear that the present experiments on electron /ion recombination in hydrogen spark channels confirm the theoretical values. Uncertainties still exist, however, in that the electron temperatures have not yet been determined for these conditions, and the general physical properties of spark channels need further investigation. The effect of radiationless 3-body collisions, leading to recombination, has been discussed by Craggs and Hopwood (1946).

§ 7. ACKNOWLEDGMENTS

The authors wish to thank Professor S. Tolansky for the use of his microphotometer, which was modified by one of us (W. H.) to a self-recorder.

The authors are also indebted to Mr. F. R. Perry for his support of the programme on spark channel investigations now being carried out in the High-Voltage Laboratory. The authors also wish to thank Sir Arthur P. M. Fleming, Director, and Mr. B. G. Churcher, Manager, of the Metropolitan-Vickers Research Department, for permission to publish this paper.

REFERENCES

Allen, H. W., 1937. *Phys. Rev.*, **52,** 1707.
Appleton, E. V., 1937. *Proc. Roy. Soc.*, A, **162,** 451.
Bates, D. R., Buckingham, R. A., Massey, H. S. W. and Unwin, J. J., 1939. *Proc. Roy. Soc.*, A, **170,** 322.
Chapman, S., 1931. *Proc. Roy. Soc.*, A, **132,** 352.
Cillié, G. C., 1932. *Mon. Not. R. Astr. Soc.*, **92,** 820.

CILLIÉ, G. C., 1936. *Mon. Not. R. Astr. Soc.*, **96**, 771.
CRAGGS, J. D., HAINE, M. E. and MEEK, J. M., 1946. *J. Instn. Elect. Engrs.* (in the press): an abstract has appeared (*J. Instn. Elect. Engrs.*, **93**, IIIA, No. 1, 191 (1946)).
CRAGGS, J. D. and HOPWOOD, W., 1946. Submitted for publication.
CRAGGS, J. D. and MEEK, J. M., 1945. *Nature, Lond.*, **156**, 21.
CRAGGS, J. D. and MEEK, J. M., 1946. *Proc. Roy. Soc.*, A., **186**, 241.
GASSIOT COMMITTEE, 1942–43. *Rep. Progr. Phys.* (*Phys. Soc.*).
GEIST, H., 1941. *Hochfreq. u. Elektroakust*, **57**, 75.
HEALEY, R. H. and REED, J. W., 1941. *The Behaviour of Slow Electrons in Gases* (Amalgamated Wireless (Australasia) Ltd.).
HOUSTOUN, R. A., 1936. *Proc. Roy. Soc., Edin.*, **57**, 163.
KENTY, C., 1928. *Phys. Rev.*, 32, 624.
LOEB, L. B., 1939. *Fundamental Processes of Electrical Discharge in Gases* (London: John Wiley).
MARGENAU, H., MCMILLAN, F. L., DEARNLEY, I. H., PEARSALL, C. S. and MONTGOMERY, C. G., 1946. *Phys. Rev.*, **70**, 349.
MASSEY, H. S. W., 1938. *Negative Ions* (Cambridge: The University Press).
MASSEY, H. S. W. and BATES, D. R., 1943. *Rep. Prog. Phys.* (*Phys. Soc.*), 9, 62.
MEEK, J. M. and CRAGGS, J. D., 1943. *Nature, Lond.*, **152**, 538.
MENZEL, D. H. and CILLIÉ, G. C., 1937. *Astrophys. J.*, **85**, 88.
MOHLER, F. L., 1928. *Phys. Rev.*, **31**, 87.
MOHLER, F. L., 1937. *J. Res. Nat. Bur. Stds., Wash.*, **19**, 559.
MOHLER, F. L. and BOECKNER, C., 1929. *J. Res. Nat. Bur. Stds., Wash.*, **2**, 489.
PENNING, F. M., 1938. *Physica*, **5**, 286.
RAYLEIGH, Lord, 1944. *Proc. Roy. Soc.*, A, **183**, 26.
SAYERS, J., 1938. *Proc. Roy. Soc.*, A, **169**, 83.
SAYERS, J., 1943. *Rep. Prog. Phys.* (*Phys. Soc.*), **9**, 52.
STUECKELBERG, E. C. G. and MORSE, P. M., 1930. *Phys. Rev.*, **36**, 16.
ZANSTRA, H., 1927. *Astrophys. J.*, **65**, 50.
ZANSTRA, H., 1946. *Proc. Roy. Soc.*, A, **186**, 236.
ZWORYKIN, V. K., MORTON, G. A. and MALTER, L., 1936. *Proc. Inst. Radio Engrs.*, **24**, 351.

THE VARIATION OF THE REFLECTIVITY OF NICKEL WITH TEMPERATURE

BY ROBERT WEIL,
South-West Essex Technical College

MS. received 16 *December* 1946; *in revised form* 18 *February* 1947

ABSTRACT. A new method for the measurement of the reflectivity of metals in the visible part of the spectrum is described: the multiple-reflection arrangement is applied to an investigation of the reflectivity of nickel at different temperatures. A novel way of polishing mirrors is given, and the construction of a vacuum furnace described. This is followed by an account of the experimental technique. Results are presented and it is suggested that the observed positive temperature coefficient of reflectivity is a relaxation phenomenon.

§ 1. METHOD

THE objects of reflectivity measurements on metals are twofold: by means of Kirchhoff's law,

$$E+R=1, \qquad \ldots\ldots(1)$$

where E stands for emissivity and R for reflectivity, we obtain a value for the former by determining the latter. If the temperature variation is known for one of these quantities this knowledge is of use in pyrometry. Also any irregularities which appear when R is plotted against the wave-length give an indication of absorption bands and thus of the natural frequencies of vibration of the bound electrons in the metal of which the mirror under investigation is made.

The use of a multiple-reflection method reduces the error of the values to small proportions and is to be preferred to the direct measurement of emissivity. Let the measured ratio of the intensity of a beam of light after n reflections to the original be S; let the error of the measurement be x, which may be positive or negative; let R be the true reflectivity of the metal; then

$$S+x=R^n, \quad S^{1/n}(1+x/S)^{1/n}=R,$$

or

$$R=S^{1/n}\left(1+\frac{x}{nS}\right). \qquad \ldots\ldots(2)$$

For a given error x the corresponding error in R will be proportional to the nth part of x. To measure small effects like the temperature coefficient of reflectivity in the visible, a multiple-reflection method thus seems indicated. Hagen and Rubens (1910) used a multiple-reflection arrangement for such a purpose, but it cannot give absolute values. The method described below has been devised in order to gather information about the temperature coefficient of reflectivity of both solid and liquid metals.

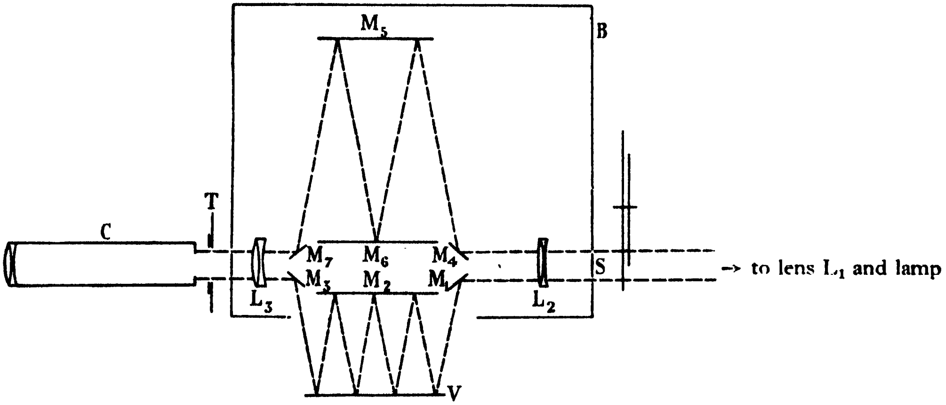


Figure 1. Elevation.

The set-up is such that four reflections take place on the surface under test, but this number can be increased. A null method is used to determine the loss in intensity due to these reflections: a beam from the same source of light as the one reflected at the mirror is compared directly with the latter, and a rotating sector is adjusted till a match is obtained between the two beams.

The source of light used was a 100-watt Mazda filament lamp run off a 36-volt battery. It was placed at the focus of a convex lens L_1 ($f=15$ cm.), and the beam uniformly illuminated a metal screen S in an aperture of the box B (figure 1). Two circular holes of equal diameter (<1 cm.) were drilled above each other in the screen S. The latter was at a distance of 5 yards from the source: this ensured that the images of the source on the mirrors mentioned below were small. The two holes were the effective source of light. A rotating sector was placed between them and the lamp. Light from the lower hole passed through an achromatic lens L_2 and was reflected from an aluminized mirror M_1 on to the specimen V. This reflected it to the mirror M_2, etc. until, after four reflections, the beam struck the mirror M_3 which reflected it through another lens L_3 and a shutter T on to the slit of the collimator C.

Clearly, the beam reflected at the specimen was also reflected at M_1 and M_3, and three times at M_2. To balance the loss in intensity due to these reflections, the other beam—the standard—must be reflected as many times under similiar conditions. This was achieved by $M_{4,5,6,7}$ as shown in figure 1. These mirrors were prepared under identical conditions: the identity of the reflecting powers of $M_{1,3,4,7}$ on the one hand, and $M_{2,5,6}$ on the other, is essential for the success of the experiments. M_7, then, reflected the standard beam through the same lens L_3 on to the slit of C just above the image of the lower hole in S. The images of the two holes touched each other for reasons which will be given below.

The mirrors M_2, M_5, M_6 and V were all sufficiently parallel: symmetry for both beams was also obtained in connection with $M_{1,3,4,7}$. The mirrors could be adjusted in several ways. Inside the box B there were six vertical iron rods, three pairs being aligned as in figure 2. Iron bosses were used to attach to them horizontal bars, and these held the frames which, in turn, gripped the mirrors (figure 3). Two screws in the small, and three in the large, frames allowed a fine adjustment to be made in the inclination of the mirrors. This was not very critical from the optical point of view, since it has been found that at an angle of incidence of 45° the reflectivity differed from that at normal incidence only by 1%, but was vital for the geometry of the arrangement. The path-lengths of the two beams focused on C had to be equal. This was achieved by fixing the length of the lower beam and adjusting the mirrors until measurement by means of a ruler indicated approximately equal paths. The final fine adjustment was carried out by fixing a thin wire across the two holes in S and adjusting the mirrors of the upper beam until the images of both holes were in focus on the slit. Then the wire was removed.

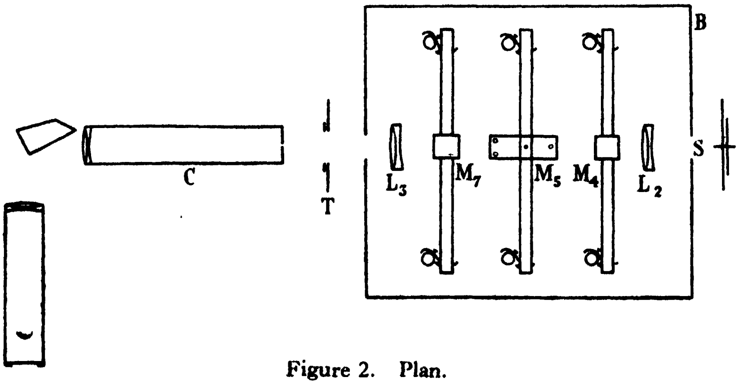


Figure 2. Plan.

On passing through the collimator, the beams were deflected by a constant deviation prism. They were then transmitted as two separate spectra through a lens which focused them on a ground-glass screen or a photographic plate.

Some further adjustment of the inclinations of the mirrors proved necessary to place corresponding wave-length ranges vertically above each other. This was ensured by replacing the filament lamp with a 250-watt Mazda mercury-vapour lamp. Both lamps were fixed to one stand, and by a simple rotation either the one or the other was made to illuminate S. Two well defined line spectra appeared on the ground-glass screen and were soon aligned. Their combined width was about 1 cm., and they were made to appear near the top of the screen: they could illuminate any part of the latter, since it could be moved vertically upward. This condition was important when photographic detection was used.

The following procedure was adopted in determining the spectral reflectivity of any specimen. The rotating double sector, devised by the author (Weil, 1947), was situated so that the standard beam was interrupted by the variable angle α_1, and the reflected beam by the fixed angle α_2. The former was varied until a match was obtained between corresponding spectral regions. If α_1 and α_2 are the two sector apertures, then

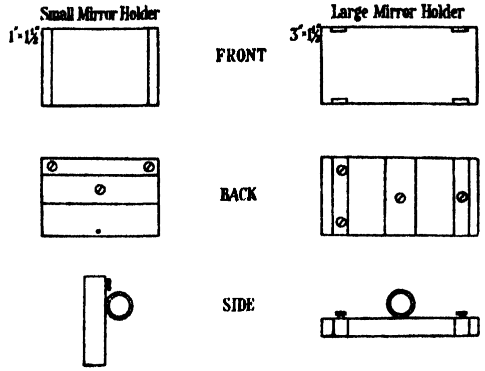


Figure 3. (Drawn to scale.)

$$R^4 = \alpha_1/\alpha_2. \qquad \ldots\ldots(3)$$

In visual photometry it is important to bear in mind the Stiles-Crawford (1933) effect. To ensure that the matching of the intensities of the two visual fields was correct, the observer was placed in such a position as to keep his head vertical and his eyes in a plane perpendicular to the ground-glass screen. This position was made comfortable. Further, the field of observation was restricted by means of a slit in a movable opaque screen to a comparatively small area on either side of the dividing line between the two spectra. The slit also limited the wave-length range under observation. When a match was obtained, the line separating the two spectra from each other could not be discerned as they were touching. The calibration of the spectra was carried out with the mercury-vapour lamp.

Several assumptions have been made in the above, and they require some scrutiny. First, it was taken for granted, as explained above, that the illumination of the screen S was uniform. Secondly, the auxiliary mirrors are supposed to have identical reflecting powers. Prepared by evaporation under identical conditions, they were opaque to visible light, and, while there is no doubt that they aged, there is no reason to expect that the rates of ageing for the various mirrors differed.

Treating $M_{1,8}$ and $M_{4,7}$ as two items, it will be seen that 120 separate experiments would have to be performed to eliminate the effect of these mirrors. Even then it is assumed that the same part of each mirror is used every time. Preliminary experiments on mercury showed that interchanging the mirrors did not affect the results.

When the specimen is enclosed in a furnace, as in the case of nickel, and the lower beam passes eight times through a silica window, the symmetry of the arrangement is destroyed unless the standard beam is reduced in intensity to a corresponding degree. This was achieved by fitting two pieces of plane transparent silica underneath M_5.

The advantages of the above method can be summed up as follows:—

1. Multiple reflection gives rise to an increased accuracy in the value obtained for the reflectivity and facilitates the detection of small effects.

2. The direct and the reflected beams are observed simultaneously: thus a change in the intensity of the source of light has no influence on the measurements.

If photographic detection is used—

3. The limit to the number of wave-lengths examined is given only by the width of the plate and the nature of the emulsion, and the record of the experiment is permanent.

4. Only one specimen mirror is required (cf. Hagen and Rubens (1910) and Burger and van Milaan (1939), in whose arrangements two are used).

5. Assuming that differences in intensity equal to 1% can be detected visually (Helmholtz puts the limit at 1/167), the accuracy is about 1% for very low reflectivities, but improves rapidly with an increase of the latter. The above figures are conservative estimates.

The method described in the previous paragraphs was tested by measuring the reflectivity of aluminium. One of the mirrors used in the present work was prepared by Messrs. A. Hilger and another by Messrs. Bellingham & Stanley. The same mirrors were investigated by an arrangement due to Bor (1937). With this apparatus another observer obtained results which agree with the author's to within the experimental error. This represents an agreement between two entirely different methods as applied to one specimen which has so far not been obtained elsewhere. Tate (1912) determined the reflectivity of steel by a catoptric and a direct method and found that the agreement between the two was poor. Evidently better agreement can be achieved by using a more accurate direct method.

§ 2. THE EXPERIMENTS ON NICKEL

The work performed by Hurst (1937), Reid (1941) and others on the infra-red emissivity and reflectivity of nickel at different temperatures led to the belief that the temperature coefficient of reflectivity would be positive in the visible parts of the spectrum. This coefficient is determined by that of the electrical conductivity when R is measured at wave-lengths $\lambda > 10\mu$, since the Hagen-Rubens formula is then valid:

$$R = 1 - 2\sqrt{\frac{c}{\lambda\sigma}}, \qquad \ldots\ldots(4)$$

where λ is the wave-length of the radiation and σ is the conductivity of the metal of which the mirror is made: c is the velocity of light. From this it is seen that, in the infra-red, the temperature coefficient of reflectivity is negative. Hence, if its positive sign can be confirmed for the visible, there must be a wave-length in the near infra-red for which the temperature coefficient is zero. The only determination of the temperature coefficient in the visible in the case of nickel has been carried out by Bidwell (1914), who measured the emissivity at a high temperature. He used the results of Hagen and Rubens (1902) for room temperature to deduce the temperature coefficient of emissivity for $\lambda = {\cdot}66\mu$: it was $-1{\cdot}25 \times 10^{-4}$ per ° c., or by (1) the corresponding reflectivity coefficient was positive. It is doubtful whether results obtained by entirely different methods can be combined in such a manner. As no measurements of the reflectivity of nickel in the visible have been carried out at different temperatures, it seemed worth while performing the work, especially as one and the same method was going to be used over a temperature range of over 400° c.

(1) *The mirrors*

Since preliminary experiments had made it desirable that the four reflections at the mirror under test should be in one line, the dimensions of the specimen were chosen as $4'' \times 0{\cdot}75'' \times 0{\cdot}2''$. Cold-rolled blocks of nickel of this size were obtained from Messrs. H. Wiggins, Birmingham. As the mirror was to be subjected to considerable temperature changes, there was no question of sputtering or evaporating nickel on glass or another substrate: the different coefficients of expansion would have ruined the surface. The only possibility, therefore, was polishing solid nickel.

Hothersall and Hammond (1940) state that, in the course of polishing nickel anodically, it has not yet been possible to eliminate the appearance of furrows: these are due to the explosive liberation of bubbles of gas on the mirror. Neither stirring nor changing the position of the anode removed this obnoxious effect in the present work. Thus a mechanical method of polishing had to be employed. Owing to the large surface of the mirror, and the occasional necessity of carrying out local polishing, it was not feasible to use a polishing machine. The whole preparation of the mirrors, with the exception of the surface-grinding, was therefore carried out manually.

To begin with, the specimen was surface-ground. The stone left marks parallel to the longest edge. After this, Oakey's emery paper was used in the usual manner. It was observed that, when the finer grades were moistened with water, the process was considerably accelerated. However, the paper dried fairly quickly and, while paraffin-oil also dried rapidly, it modified the surface of the paper in such a way as to render unnecessary its repeated application. Polishing the surface immediately afterwards with a paste of levigated magnesia powder in distilled water on Selvyt cloth gave rise to polishing pits. According to Ferguson (1943), these can be reduced by using No. 100 carborundum in a 1 : 1 liquid soap solution on a paraffin disc as an intermediate stage. The writer found it possible to eliminate them altogether by using a pitch surface in the following manner. Commercial pitch was heated until liquid, when it was poured on to a flat surface. A piece of plane glass, on which moist soap had been rubbed to prevent the pitch from sticking to it, was pressed on the latter when it was about to solidify. This surface was not perfectly continuous, but the small crevices and hollows in it were useful as reservoirs of the thin suspension of alumina in distilled water, then poured on it. After a few minutes' polishing, striking results were obtained: the 4/0 scratches had more or less disappeared. The pitch surface was washed and dried, as was the specimen, and distilled water poured on the former. The polishing process was continued, the direction of the movement being changed, as the alumina was found to have a slight abrasive action. Finally, a piece of Selvyt cloth was wrapped round a small wooden block, a paste of magnesia applied, and the mirror touched up. It was possible to obtain very good mirrors consistently. As comparison with other work showed, their room-temperature reflectivity was amongst the highest measured. This is of interest, since Coblentz and Stair (1929) also used a solid nickel mirror made of a rolled rod of the refined metal, but their results are rather low.

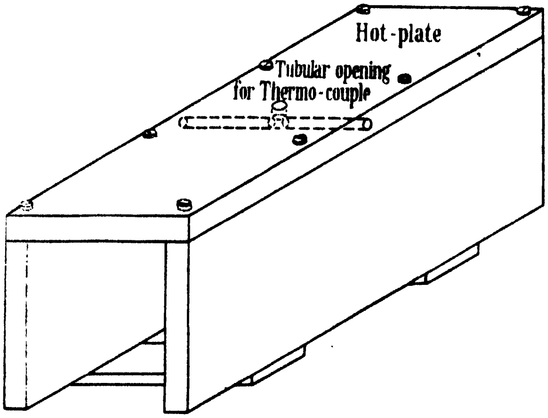


Figure 4.

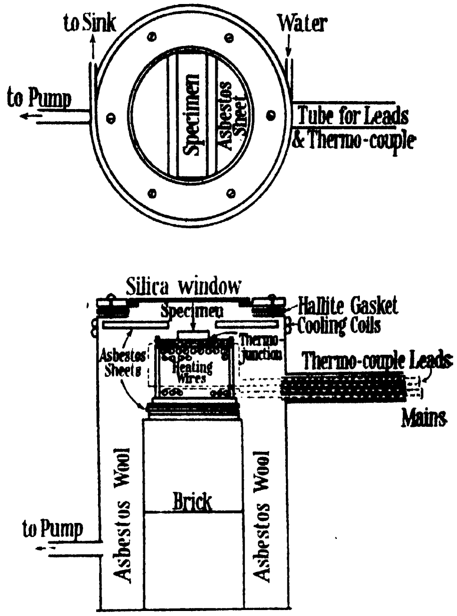


Figure 5.

(2) *The furnace*

After some preliminary work it was decided to use resistance heating. The specimen was placed on a hot-plate which formed part of the heating element. The latter consisted of a box (figure 4). Two steel plates, (5·75″ × 1·5″ × $\frac{1}{8}$″) were connected by two strips of steel along one of their longitudinal edges, and by the hot-plate (5·75″ × 1·5″ × 0·25″) along the other. A tubular opening, 0·1″ in diameter, was drilled into the hot-plate parallel to its large planes and at right angles to its longitudinal edge, and a hole, $\frac{3}{16}$″ in diameter, at its centre. One small silica tube was placed into the opening on either side of the hole, and a platinum platinum-rhodium thermocouple threaded through them. The spot-welded junction protruded through the hole. Enough spring was provided for the specimen to depress the junction when lying on the plate: thus contact was ensured between the thermocouple and the specimen.

Sixty silica tubes, whose length was 6″, external diameter 3 mm. and bore 1 mm., were threaded on more than ten yards of Brightray wire (S.G.30). This chain was folded up, care being taken to avoid potential short circuits between the joints, and placed in the framework described above.

The furnace proper consisted of a cylinder (figure 5). A thin sheet of steel was bent into a cylindrical shape, the two straight edges being silver-soldered together. To one end of this cylinder, whose height was 10″ and whose diameter 7″, there was silver-soldered a circular plate made of mild steel, and to the other a circular flange. The external diameter of the latter was 7″, its internal 5·5″. The frame carrying the quartz window was screwed to it. It was made of another piece of mild steel, diameter 7″, thickness ⅜″. An aperture of 4·5″ was cut into it, and a recess of 5·25″ was cut out of the plate: the quartz-window, whose diameter was 5″, fitted into it, an allowance being made for the vacuum wax. The window was fixed to the frame by pouring liquid Apiezon W (melting point 80–90° c.) into the recess, placing the silica plate there, and playing the flame of a Mekker burner around the frame until all the air holes in the wax had disappeared.

A steel tube of 3″ length and about 0·5″ diameter was fitted near the top of the furnace. The heater and thermocouple wires, adequately insulated by means of silica beads, were threaded through it, and connected to two pairs of terminals on two ebonite blocks. The latter were fixed rigidly to the tube. Some insulating tape was wound round the open end of the tube, and provided a suitable base on which to deposit vacuum wax. Diametrically opposite to this tube, another opening was made near the bottom of the furnace, and the connection fitted to a vacuum pump.

Water running through two turns of copper tubing cooled the joint between the furnace and the window-frame to such an extent that the vacuum wax applied there stayed solid even when the temperature in the furnace was 500° c. or more.

Two half-bricks were placed on top of each other inside the furnace. Several layers of asbestos board were put on top of them, and the heating element rested on the latter. Its direction was perpendicular to the line joining the two tubes mentioned above. No radiation therefore reached them directly, and thus they stayed cool. The whole of the furnace was lagged with asbestos wool. Two semi-circular discs of asbestos were placed just below the silica window, leaving an adequate gap for the beam reflected to and fro between the specimen and M_2.

(3) *Experimental technique*

Since it was not desirable to confine the measurements to one mirror only, several were prepared, or, alternatively, when one had been examined its surface was ground down with aluminium oxide and a fresh one prepared. The technique involved in carrying out these measurements will be described in the following.

The freshly polished mirror was placed on the hot-plate. Copper-foil was put around it to act as a getter. The pressure in the furnace could be reduced with a Cenco-Hyvac pump to 1 or 2 mm. of mercury and did not exceed 4 mm. at the highest temperatures used. The presence of the getter ensured, however, that the residual oxygen did not tarnish the specimen. This was easily ascertained by measuring the reflectivity also when the mirror cooled: since these results tallied for identical temperatures with those obtained when the mirror was being heated, it was concluded that no noticeable tarnishing took place. Then a Hallite gasket was placed on the flange, and the window-frame screwed to the furnace by means of six screws. Apiezon W wax was next applied to the joint and a Mekker flame played over it. The six screw-heads were similarly covered: this operation lasted about five to seven minutes. Next the pump was started to test the pressure and to detect any leaks. Then the furnace was placed on a wooden box below M_2. The mains were connected to the heater, and a potentiometer, made by Messrs. Tinsley & Co., Type 3387 B, to the thermocouple. The water-mains and sink were fitted to the cooling coil. When the pressure in the furnace was a minimum, the mirrors $N_{2,3}$ were adjusted so that the reflected beam fell on the slit of the spectroscope as previously described. When this was done, and the equality of the path-lengths of the two beams ensured, the reflectivity of nickel was determined for different wave-lengths at room temperature. Then the heater, controlled by a transformer and a rheostat, was switched on. The reflectivity was measured at intervals of 80 or 100° C. and checked again as the furnace cooled.

(4). *Results*

The experimental results are presented in the table below. The temperature coefficient is defined by

$$\alpha = \frac{1}{R}\cdot\frac{dR}{dT}. \qquad \text{......(5)}$$

A priori it is possible for dR/dT to be less affected by the condition of the surface of the specimen than α. This surmise was confirmed: initially, when not much experience had been gained in polishing mirrors, R was low, but dR/dT did not differ appreciably from later values. This is the reason why the results for the gradient are quoted in addition to those for R and α. The data represent mean values obtained from a number of measurements. Plotting R against the temperature gives straight lines for the whole range in which measurements were taken.

Table

Wave-length (μ)	0·47	0·53	0·59	0·64
Reflectivity (%) at room temperature..	58·0	62·8	65·5	66·8
Reflectivity gradient, $\frac{dR}{dT}$ ($\times 10^3$) ..	9·28	7·53	6·55	5·68
Reflectivity coefficient, α ($\times 10^4$) ..	1·61	1·2	1·01	0·85

It is to be noted that, owing to a slight temperature gradient across the specimen, the temperature of the mirror was somewhat less than that given by the thermocouple. This is of no consequence as far as the results are concerned. The accuracy of the values varied from 0·9% in the blue to 0·5% in the red regions of the spectrum, improving with the reflectivity, as mentioned before. There was no discontinuity at the Curie point. This was to be expected since Löwe (1936) has shown in some experiments on the emissivity of nickel in the infra-red that such a discontinuity does not occur for wave-lengths $\lambda < 4{\cdot}5\ \mu$. Ornstein and Kofoed (1938), however, find a decrease in R amounting to 0·5% for a wave-length of 0·65 μ. Nor is there a discontinuity at what Schubert (1937) calls the temperature of recrystallization, approximately 320° c. This refers only to the recrystallization of the amorphous surface layer of the mirror, for it is well known that the whole of the metal crystallizes at temperatures much nearer to the melting point. The values of α are plotted against the wave-length in figure 6. This makes it possible to compare them with those obtained by Hurst (1933) (emissivity) and Reid (1941) (reflectivity). The latter quotes in his results only three values, one of them at 5 μ, and consequently a broken line represents his work.

It is seen that the temperature coefficient of the reflectivity of nickel in the visible parts of the spectrum is positive. Its numerical value increases as the violet end of the spectrum is approached, but this is largely due to the fact that the reflectivity is reduced. The writer (Weil, 1946) suggested a possible explanation of the positive sign of this coefficient. In the far infra-red the coefficient is negative since, as is seen from (4), it depends only on the temperature coefficient of the conductivity. As the visible parts of the spectrum are approached, the dielectric portion of the metal plays a greater rôle and it can be shown that its temperature coefficient tends to diminish the absolute value of dR/dT. If the term for the dielectric is greater than the one representing the free electrons, a positive value is obtained for α. The fact that the time of relaxation of the free electrons should be considered in the short-wave region is also relevant. It can be shown that under such conditions the temperature coefficient of $\sigma(\nu)$ (the conductivity for a frequency ν) is positive. An analysis of the whole problem will be presented in the near future.

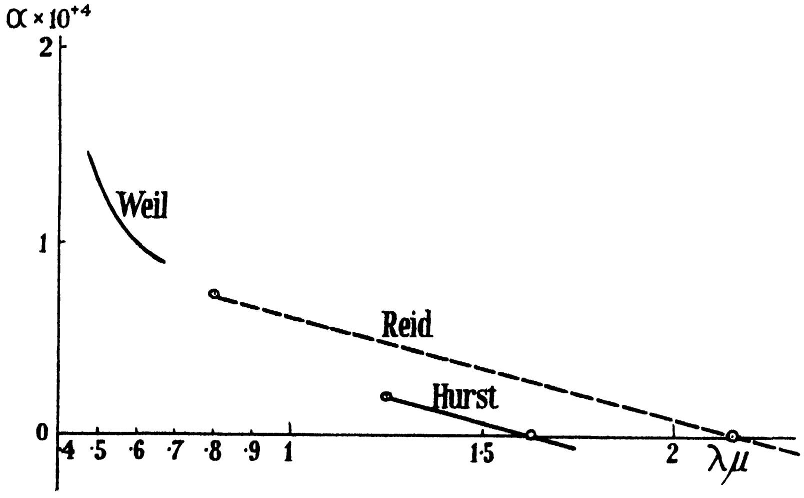


Figure 6.

ACKNOWLEDGMENTS

The above work has been carried out under the auspices of the Pyrometry Sub-Committee, British Iron and Steel Research Association. The writer would like to thank Dr. H. Lowery, for his interest in the work, and Mr. J. Bor, M.Sc. (Tech.), for many valuable discussions. The help given by Mr. R. J. Donato, B.Sc., and Mr. W. Bennett is also greatly appreciated.

REFERENCES

BIDWELL, 1914. *Phys. Rev.*, **3**, 439.
BOR, 1937. *Nature, Lond.*, **139**, 716.
BURGER and VAN MILAAN, 1939. *Physica*, **6**, 435.
COBLENTZ and STAIR, 1929. *Bur. Stand. J. Res., Wash.*, **2**, 343.
FERGUSON, 1943. *Metal Progress*, **43**, 743.
HAGEN and RUBENS, 1902. *Ann. Phys., Lpz.*, **8**, 1; 1910. *S.B. Preuss. Akad. Wiss.*, **23**, 467.
HOTHERSALL and HAMMOND, 1940. *J. Electrodepositors' Technical Soc.*, **16**, 83.
HURST, 1933. *Proc. Roy. Soc.*, A, **144**, 466.
LÖWE, 1936. *Ann. Phys., Lpz.*, **25**, 212.
ORNSTEIN and KOFOED, 1938. *Physica*, **5**, 175.
REID, 1941. *Phys. Rev.*, **60**, 161.
SCHUBERT, 1937. *Ann. Phys., Lpz.*, **29**, 473.
STILES and CRAWFORD, 1933. *Proc. Roy. Soc.*, B, **112**, 428.
TATE, 1912. *Phys. Rev.*, **34**, 321.
WEIL, 1946. *Nature, Lond.*, **158**, 672; 1947. *Proc. Phys. Soc.*, **59**, 161.

MAGNETIC FOCUSING BETWEEN INCLINED PLANE POLE-FACES*

BY H. O. W. RICHARDSON,
George Holt Physics Laboratory, University of Liverpool
(Now at Department of Natural Philosophy, University of Edinburgh)

MS. received 20 *November* 1946

ABSTRACT. Electron orbits have been computed in the magnetic field between inclined equipotential plane pole-faces. The field, if applied to β-ray spectroscopy, should give high dispersion and resolving power combined with a fair solid angle of collection Ω. In a particular case, Ω is estimated to be 16·5 times that of a conventional semi-circular spectrograph of similar dimensions.

If the inclined planes are bevelled so as to become parallel at their closest parts, both lateral and longitudinal focusing should occur. This case is treated approximately, and it seems probable that, although aberrations may reduce the resolution below that given by the first method, high values of Ω (of the order of 1 per cent of 4π) may be obtainable.

§ 1. INTRODUCTION

METHODS have recently been described for increasing the resolving power of the type of β-ray spectrograph which uses semi-circular focusing (Voges and Ruthemann, 1939; Korsunsky, Kelman and Petrov, 1945). The increase has been obtained by making small departures from uniformity in the magnetic field, for example—by altering the profile of the pole-faces of the magnet. In order to get good reproducibility and ease of construction, the simplest possible profile is evidently desirable.

It is thus of interest to calculate the orbits of electrons in the non-uniform field between two inclined plane pole-faces which are treated as infinite equipotential planes.

---

* Apart from minor corrections, these calculations were given in a report circulated in January 1944, but publication was delayed by the war.

The field is the same as that due to a straight current flowing along the line of intersection of the two planes, and the lines of force are arcs of circles.

The orbits lose their circular shape and become periodic loops which are a special case of the "enroulements trochoïdaux" used by Thibaud and others (Thibaud, 1938). Some features of the orbits, such as their turning points, have been given by J. J. Thomson (1903).

It seems possible that a marked improvement in the resolving power and the solid angle of collection could be obtained in a spectrograph in which the orbits traversed a region in which the field strength varied by a factor of about 10. The gain in solid angle is partly due to the rate of sideways spreading of the beam being less than it is in the uniform field.

The sideways spreading can be further reduced by a modification of the profile of the pole-faces. In the modified profile, shown in figure 5, the inclined planes are bevelled so as to become parallel at their closest parts. The orbits in the case of bevelled poles are treated only approximately, but it seems that this field should give some degree of both lateral and longitudinal focusing and should avoid the attenuation of the beams due to sideways spreading, which lowers the solid angle of collection of the semi-circular spectrograph.

As the pole-faces are treated as equipotentials, we assume infinite permeability and ignore the field of the magnetizing current.

§ 2. MATHEMATICAL THEORY

(a) *The magnetic vector potential*

Let r, z, and ϕ be cylindrical coordinates representing radial, longitudinal and azimuthal displacements with respect to the axis $r=0$ defined by the line of intersection of the two planes which contain the inclined pole-faces. The equipotential planes $\phi=$ constant all intersect along this axis, on which lie the centres of the circular lines of force. The field $\vec{H}$ satisfies

$$H=H_1/r, \qquad \ldots\ldots(1)$$

where H_1 is the intensity at $r=1$. The vector potential $\vec{A}$ obeys $\vec{H}=\operatorname{curl}\vec{A}$ and, by Stokes' theorem,

$$\int(\vec{H}\,.\,\vec{d\sigma}) = \oint(\vec{A}\,.\,\vec{ds}), \qquad \ldots\ldots(2)$$

where $\vec{ds}$ is a vector forming an element of a closed path and $\vec{d\sigma}$ is a vector normal to an element of area $d\sigma$ of the surface bounded by the path over which the integral is taken. In our case $\vec{A}$ reduces to a single component parallel to the axis $r=0$, varying only with r, and of magnitude $A(r)$. Apply (2) to a rectangular path 1234 lying in an equipotential plane, then

$$\iint H_1/r\,.\,dr\,dz = A(r_1)(z_2-z_1)+A(r_2)(z_4-z_3). \qquad \ldots\ldots(3)$$

Thus $\quad A(r_1)-A(r_2)=H_1\log r_1/r_2. \qquad \ldots\ldots(4)$

Let $A=0$ when $r=0$. Then

$$A(r)=H_1(1-\log r). \qquad \ldots\ldots(5)$$

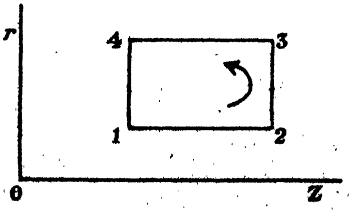


Figure 1.

(b) *Three-dimensional equations of the orbits*

The relativistic Lagrangian for a particle of rest-mass m and charge e_0, in e.m.u., moving with velocity $\vec{v}$ in a magnetic field, is

$$L = mc^2(1-\sqrt{1-v^2/c^2}) + e_0(\vec{v}\,.\,\vec{A}). \qquad \ldots\ldots(6)$$

$v^2 = \dot{r}^2 + \dot{z}^2 + r^2\dot{\phi}^2$ is a constant of the motion because the energy does not change. From (5)

$$L = mc^2(1-\sqrt{1-v^2/c^2}) + H_1 e_0 \dot{z}(1-\log r). \qquad \ldots\ldots(7)$$

The angular momentum p_ϕ about the z-axis is

$$p = \frac{\partial L}{\partial \dot{\phi}} = \frac{mr^2\dot{\phi}}{\sqrt{1-v^2/c^2}} = m_t\,.\,r^2\dot{\phi},$$

where m_t is the constant transverse mass. The first of Lagrange's equations gives $\dfrac{d}{dt}\left(\dfrac{\partial L}{\partial \dot{\phi}}\right) = \dfrac{\partial L}{\partial \phi} = 0$. Thus p_ϕ is constant. This is because the Lorentz force $\vec{F}$ always lies in an equipotential plane, so that it has no turning moment about the common axis of intersection of all these planes.

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{r}}\right) = \frac{d}{dt}(m_t\,.\,\dot{r}) = \frac{\partial L}{\partial r} = m_t\,.\,r\dot{\phi}^2 - \frac{H_1 e_0 \dot{z}}{r}. \qquad \ldots\ldots(8)$$

Thus

$$\ddot{r} = -\frac{H_1 e_0 \dot{z}}{r m_t} + \frac{p_\phi^2}{r^3 m_t^2}, \qquad \ldots\ldots(9)$$

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{z}}\right) = \frac{d}{dt}(m_t\,.\,\dot{z} + H_1 e_0(1-\log r)) = \frac{\partial L}{\partial z} = 0. \qquad \ldots\ldots(10)$$

Thus

$$\ddot{z} = \frac{H_1 e_0 \dot{r}}{m_t\,.\,r} \quad \text{and} \quad \dot{z} = \frac{H_1 e_0 \log r}{m_t} + C. \qquad \ldots\ldots(11)$$

At the longitudinal turning points of the orbits, $\dot{z} = 0$. Let the turning points be at $r = A$. Then

$$\dot{z} = \frac{H_1 e_0}{m_t}\,.\,\log\frac{r}{A}. \qquad \ldots\ldots(12)$$

Putting this in (9) and using $\int \ddot{r}\,dr = \int \dot{r}\,d\dot{r} = \dfrac{\dot{r}^2}{2} + \text{constant}$,

$$\dot{r}^2 = -\left(\frac{H_1 e_0}{m_t}\,.\,\log\frac{r}{A}\right)^2 - \frac{p_\phi^2}{r^2 m_t^2} + c'. \qquad \ldots\ldots(13)$$

From (12) and (13), $c' = \dot{r}^2 + \dot{z}^2 + r^2\dot{\phi}^2 = v^2$. Let

$$K = \frac{m_t\,.\,v}{H_1 e_0} \quad \text{and} \quad a_0 = \frac{p_\phi}{m_t\,.\,v}. \qquad \ldots\ldots(14)$$

K is proportional to the momentum $m_t\,.\,v$ if H_1 is constant.

From (13)

$$\dot{r} = \pm v\sqrt{1 - \frac{a_0^2}{r^2} - \frac{1}{K^2}\left(\log\frac{r}{A}\right)^2}. \qquad \ldots\ldots(15)$$

For computing the orbits it is convenient to express r, z, and ϕ in terms of a parameter ψ which satisfies $v\cos\psi = \dot{z}$. We see that ψ is the angle between the velocity vector $\vec{v}$ at a point P on the orbit and a vector through P parallel to the axis $r=0$, pointing in the direction of increasing z. Equation (12) becomes

$$r = A \exp . \frac{m_t . \dot{z}}{H_1 e_0} = A e^{K \cos\psi}, \qquad \ldots\ldots(16)$$

$$z = \int \frac{\dot{z}}{\dot{r}} dr = \frac{1}{K} \int \frac{\log \frac{r}{A} \cdot dr}{\sqrt{1 - \frac{a_0^2}{r^2} - \frac{1}{K^2}\left(\log \frac{r}{A}\right)^2}}$$

$$= -AK \int \frac{e^{K \cos\psi} . \cos\psi . d\psi}{\sqrt{1 - \frac{a_0^2 . e^{-2K \cos\psi}}{A^2 \sin^2\psi}}}, \qquad \ldots\ldots(17)$$

$$\phi = \int \dot{\phi}\, dt = a_0 v \int \frac{dt}{r^2}$$

$$= a_0 v \int \frac{dr}{\dot{r} r^2} = -\frac{a_0 K}{A} \int \frac{e^{-K \cos\psi} . d\psi}{\sqrt{1 - \frac{a_0^2 e^{-2K \cos\psi}}{A^2 \sin^2\psi}}}. \qquad \ldots\ldots(18)$$

Equations (12) and (15) were given by Thomson (1903), who used them to find the value of r at the turning points at which either $\dot{z}$ or $\dot{r}$ is zero.

(c) *Orbits from a point source*

Let a point source S lie in the mid-plane $\phi=0$ at $r=r_s$. At S let $\psi=\psi_s$ and let the velocity vector be inclined at an angle ξ_s to the mid-plane. ξ_s is a measure of the angular width of the emitted beam and is related to the constant a_0 by

$$a_0 = \frac{p_\phi}{m_t . v} = r \sin\xi = r_s \sin\xi_s. \qquad \ldots\ldots(19)$$

ξ is the inclination of the orbit at any point P to the meridional plane $\phi=$ constant, passing through P.

Consider a family of orbits of equal momentum coming from S with different angles of emission ψ_s. Each orbit has $r=r_s$ when $\psi=\psi_s$ so that, from (16),

$$A = r_s . e^{-K \cos\psi_s}. \qquad \ldots\ldots(20)$$

The equations of the family are

$$r = r_s . e^{K(\cos\psi - \cos\psi_s)}, \qquad \ldots\ldots(21)$$

$$z = -K r_s . e^{-K \cos\psi_s} \int \frac{\cos\psi . e^{K \cos\psi} . d\psi}{\sqrt{1 - \frac{\sin^2\xi_s}{\sin^2\psi} . e^{-2K(\cos\psi - \cos\psi_s)}}}, \qquad \ldots\ldots(22)$$

$$\phi = -K \sin\xi_s . e^{K \cos\psi_s} \int \frac{e^{-K \cos\psi} . d\psi}{\sqrt{1 - \frac{\sin^2\xi_s}{\sin^2\psi} . e^{-2K(\cos\psi - \cos\psi_s)}}}. \qquad \ldots\ldots(23)$$

The integrands in (22) and (23) become infinite when $\psi(=\psi_m)$ is one of the solutions of

$$\sin^2\psi . e^{2K \cos\psi} = \sin^2\xi_s . e^{2K \cos\psi_s}. \qquad \ldots\ldots(24)$$

The infinities occur at the maxima and minima of r, where $\dot{r}=0$. The parameter ψ is excluded from the ranges of values in which the integrand is imaginary.

These ranges are centred on $\psi = n\pi$, where n is an integer, and are small if ξ_s is small.

(d) *Orbits lying in the mid-plane*

In β-ray spectroscopy the image will be formed by orbits which lie in or near the mid-plane $\phi = 0$, and for which the inclination ξ is small or zero. For $\xi_s = 0$, the equations of the family of coplanar orbits are

$$r = r_s \, . \, e^{K(\cos\psi - \cos\psi_s)}, \qquad \ldots\ldots(25)$$

$$z = -Kr_s \, . \, e^{-K\cos\psi_s} \int \cos\psi \, . \, e^{K\cos\psi} \, . \, d\psi. \qquad \ldots\ldots(26)$$

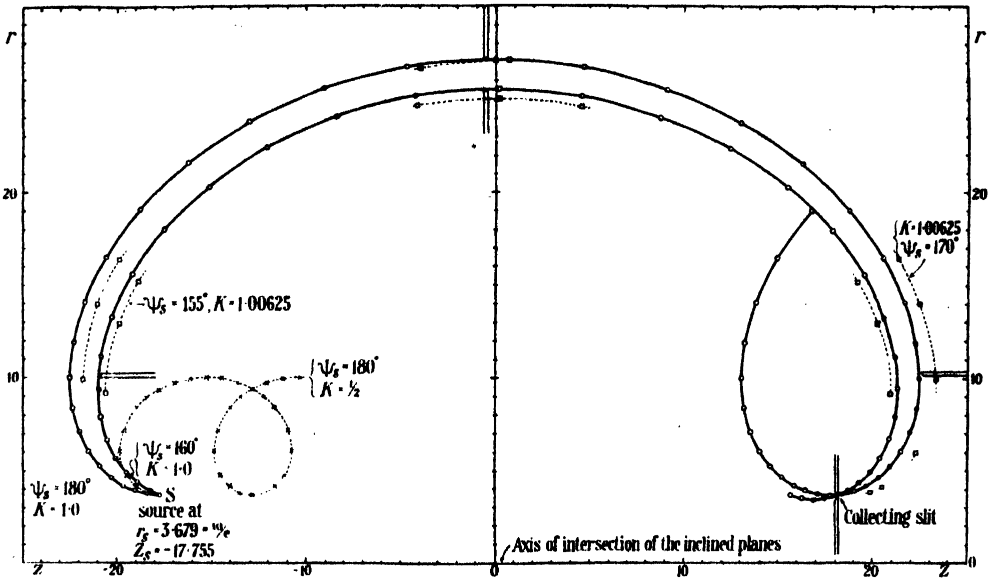


Figure 2. Orbits in the mid-plane $\phi = 0$ between inclined plane pole-faces. The baffles are adjusted for high resolution. They completely exclude the two ranges of rays with 0·625% higher momentum which would otherwise enter the slit. One narrow range is near $\psi_s = 155°$. The other, a wider one, is from 170° to about 190°.

If we put the source S at $z = 0$, the lower limit of the integral is at $\psi = \psi_s$. The integrand is always real, so ψ can now have any value. r varies periodically between Ae^{K} and Ae^{-K}, with maxima at $\psi = 2n\pi$ and minima at $(2n+1)\pi$, where n is an integer. The orbits, which are shown in figure 2, are a series of loops in which the radius of curvature ρ is equal to Kr. The z-displacement in a complete loop increases rapidly with K, as can be seen from the orbits with $K = \frac{1}{2}$ and $K = 1$. The loops contract to coincident circles as K, which is proportional to the momentum, approaches zero. Orbits with the same value of K have the same shape but different sizes, because both r and z are proportional to $e^{-K\cos\psi_s}$, which varies with the angle of emission ψ_s.

Let z_0 be the integral of (26) from $\psi = 0$ to $\psi = \psi$.

$$z_0 = -Kr_s . e^{-K\cos\psi_s}\Bigg[\frac{K\psi}{2} - \frac{K}{4} . \sin 2\psi + \sin\psi . e^{K\cos\psi}$$
$$+ \frac{K^2}{3} . \sin^3\psi . \left(1 + \frac{K^2}{3.5} + \frac{K^4}{3.5^2.7} + \frac{K^6}{3.5^2.7^2.9} + \ldots\right)$$
$$+ \frac{K^3}{16} . \left(\psi - \frac{\sin 4\psi}{4}\right)\left(1 + \frac{K^2}{4.6} + \ldots\right) + \frac{K^4}{3!5} \sin^3\psi . \cos^2\psi\left(1 + \frac{K^2}{5.7} + \ldots\right)$$
$$+ \frac{K^5}{4!6} . \sin^3 . \cos\psi . \left(1 + \frac{K^2}{6.8} + \frac{K^4}{6.8^2.10} + \ldots\right) + \text{etc.}\Bigg]. \qquad \ldots\ldots(27)$$

It is usually easier to find z by numerical integration than by summing the terms in (27). Because the loops are symmetrical about the maxima and minima it is convenient to use the central difference formula (Whittaker and Robinson, p. 146), starting at $\psi = 0$. Values of $\int_0^{\psi} e^{K\cos\psi} . \cos\psi d\psi$ are given in table 1 for $K = \frac{1}{2}$, 1·0, and 1·00625. The close pair of values can be used to estimate the resolving power.

Table 1. Values of $\int_0^{\psi} e^{K\cos\psi} . \cos\psi . d\psi$, in radians. (For $K = \frac{1}{2}$ and 1·00625, 5-figure, and for $K = 1$, 7-figure, trigonometrical functions were used.)

ψ		K		ψ		K	
(°)	0·5	1·0	1·00625	(°)	0·5	1·0	1·00625
10	0·2856	0·46965	0·4726	100	1·4755	2·23040	2·2421
20	0·5583	0·91157	0·9173	110	1·4360	2·19582	2·2076
30	0·8068	1·30243	1·3103	120	1·3765	2·14771	2·1596
40	1·0219	1·62648	1·6360	130	1·3015	2·09147	2·1035
50	1·1975	1·87685	1·8875	140	1·2150	2·03071	2·0430
60	1·3309	2·05483	2·0660	150	1·1202	1·96774	1·9804
70	1·4221	2·16785	2·1793	160	1·0197	1·90386	1·9169
80	1·4737	2·22688	2·2385	170	0·9158	1·83976	1·8531
90	1·4898	2 24395	2·2556	180	0·8102	1·77550	1·7893

The equation $r = A . e^{K\cos\psi}$ was given by Larmor (1884) for the catenary curve assumed by a flexible conductor carrying a current and lying in a plane in a magnetic field produced by a stronger current flowing along the axis $r = 0$, in the same plane.

§ 3. APPLICATION OF THEORY

(a) *Focusing of orbits lying in the mid-plane*

Because the orbits are symmetrical about the maximum at $\psi = 0$, a ray emitted from S with $\psi = \psi_s$ will return to the initial radius r_s when $\psi = -\psi_s$. On plotting the orbits, it seems that the greatest concentration of returning rays is near $r = r_s$, so we expect to find an image I of the source S at some point on this radius. Let z_i be the distance from S to the point at which the ray returns to $r = r_s$. Then from (26)

$$z_i = -2Kr_s . e^{-K\cos\psi_s}\int_{\psi_s}^{0} \cos\psi . e^{K\cos\psi} . d\psi. \qquad \ldots\ldots(29)$$

The displacement z_i is plotted against ψ_s in figure 3. It can be seen, for $K = 1$, to have a maximum at $\psi_s = 169°$, where the envelope of the family of orbits crosses $r = r_s$, giving an image with a sharp outer edge just as in semicircular focusing.

At $\psi_s = 180°$, a second stationary value of z_i can be seen. This arises because the loop near the image is cut by the line $r = r_s$ in two points which become coincident when $\psi_s = \pi$. The line is then a tangent to the loop. For certain positions of the slit this second kind of image can make a small addition to the solid angle of collection.

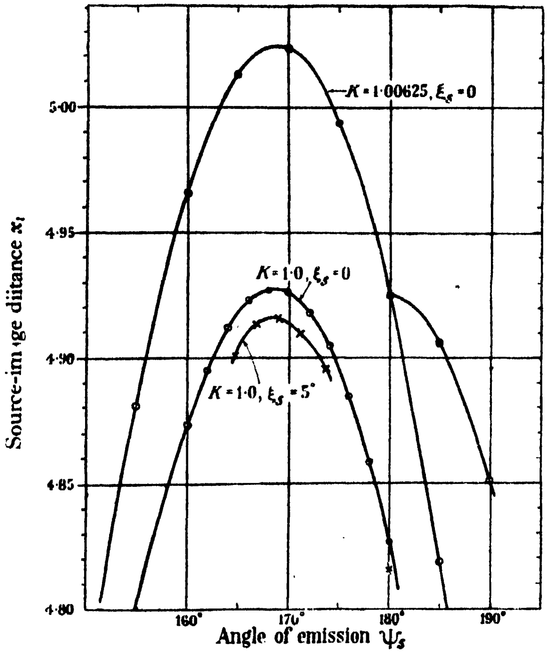


Figure 3. Focusing of orbits of differing inclinations ξ_s and momenta KH_1e_0.

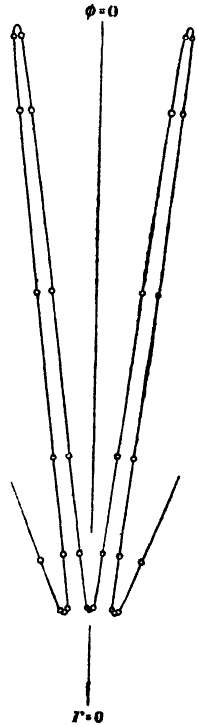


Figure 4. The projection of an orbit on the (r, ϕ) plane. Initial inclination $\xi_s = 5°$, with $\psi_s = \pi$, and $K = 1$.

(b) *Focusing of inclined orbits*

For inclined rays, (29) becomes

$$z_i = -2Kr_s \, . \, e^{-K\cos\psi_s} \int_{\psi_s}^{\psi_m} \frac{\cos\psi \, . \, e^{K\cos\psi} \, . \, d\psi}{\sqrt{1 - \dfrac{b^2}{\sin^2\psi} \cdot e^{-2K\cos\psi}}}, \qquad \ldots\ldots(30)$$

where

$$b = \sin\xi_s \, . \, e^{K\cos\psi_s}. \qquad \ldots\ldots(31)$$

The upper limit ψ_m is a value of ψ satisfying (24), at which the integrand is infinite. In the particular case in which $\psi_s = \pi - \xi_s$, the lower limit ψ_s also satisfies (24). The three lines forming the angles ψ_s and ξ_s are then coplanar, so that the projection on the mid-plane at S of $\vec{v}$ is parallel to $r = 0$. Thus S is at a minimum of r and an integration with the limits $\pi - \xi_s$ and ψ_m covers a complete half-cycle of ψ.

Expressions (23) for ϕ and (30) for z_i have been integrated numerically over this range of ψ with $\xi_s = 5°$. The projection of the orbit in the (r, ϕ) plane is plotted in figure 4. It consists of periodic peaks and troughs radiating from $r = 0$.

The angle ϕ_0 between adjacent troughs is 14°·60. For other values of ξ_s it is nearly proportional to $\sin \xi_s . e^{-K \cos \xi_s}$ because the integral in (23) over a complete cycle of ψ only varies slowly with ξ_s for small values of ξ_s. It increases from 7·6647 to 7·9204 radians between $\xi_s = 0$ and 5°.

The integrations are made difficult by the infinities at the termini of ψ, which make successive differences Δ decrease slowly. ψ was broken up into the five ranges $175°-170°-160°-10°-2°-\psi_m = 40{\cdot}298$ minutes. The three inner ranges were integrated by the Gregory formula using intervals h of 2°, 10° and 1° respectively. For the outer ranges the formula of Jeffreys (1939) was used:

$$\int_0^{3h} (\alpha x^{-\frac{1}{2}} + \beta x^{\frac{1}{2}} + \gamma x^{\frac{3}{2}}) dx = h\left(\frac{14}{5}\sqrt{3} . y_1 - \frac{8}{5}\sqrt{6} . y_2 + \frac{12}{5} y_3\right). \quad \ldots\ldots(32)$$

y_1, y_2, and y_3 are the values of the integrand at $x = h$, $2h$, and $3h$ respectively.

In order to locate the image formed by rays with the same inclination ξ_s, it is necessary to find z_i with different angles of emission ψ_s. This introduces a difficulty because b in (30) then varies with ψ_s and a new set of integrands must be computed for each value of ψ_s. We can avoid this labour by allowing ξ_s to vary with ψ_s by the small amount needed to keep b constant. Then the same set of integrands can be used for all values of ψ_s, with the same upper limit at $\psi = \psi_m$.

Let $\sin \xi_s = b . e^{-K \cos \psi_s}$, where $b = \sin 5° . e^{-\cos 5°}$. Using $K = 1$ and $r_s = 1$, we get the values of z_i in table 2, which also contains some values for comparison found for rays in the mid-plane, with $\xi_s = 0$. In figure 3, z_i is plotted against ψ_s' where ψ_s' is the angle between the axis $r = 0$ and the projection on the mid-plane of the velocity vector $\vec{v}$ at S. It satisfies the relation

$$\cos \psi_s = \cos \psi_s' . \cos \xi_s. \quad \ldots\ldots(33)$$

The use of the projected angle ψ' for inclined rays is convenient because, unlike ψ, it is excluded from no ranges of values.

Table 2

ψ_s	ψ_s'	ξ_s	z_i	ψ_s	ξ_s	z_i
175°	180°	5°	4·8156	180°	0	4·8263
172°	173° 45′	4° 58′	4·8953	172°	0	4·9178
170°	171° 6′	4° 57′	4·9096	170°	0	4·9261
168°	169° 2′	4° 55′	4·9153	168°	0	4·9269
166°	166° 52′	4° 52′	4·9135	166°	0	4·9222
164°	164° 42′	4° 50′	4·9049	164°	0	4·9115

The focusing of the inclined rays closely resembles that of the rays in the mid-plane, but the maximum displacement z_i is less by a factor of 0·9976. This shortening can be seen from the curve for $K = 1{\cdot}00625$ to be equivalent to that due to a decrease in momentum of 0·08%. It is less than with semi-circular focusing, where the shortening factor is $\cos \xi_s = 0{\cdot}99620$.

The length l of the image will be approximately $2r_s . \phi_0$ because most of the rays which form the image will have ψ_s' just less than π, so that they will return to the initial radius after an azimuthal displacement ϕ just less than that between two troughs, ϕ_0. With the semi-circular spectrograph the image length l' is

$2\pi\rho \sin \xi_s$. With inclined poles with the same ξ_s, and $z_i = 2\rho$, the image length $2r_s\phi_0$ is only 0·3785 l'. This shortening of the image is due to the lower sideways spreading of the beam caused by the presence over the long outer parts of the inclined orbits of an inward component of the Lorentz force directed back towards the mid-plane. In the uniform field, in contrast, the rays spiral away from the mid-plane with the constant outward velocity $v \sin \xi_s$.

The curvature of the image is a little greater than with semi-circular focusing, because the smaller shortening of z_i is outweighed by the larger shortening of the image length l. The values of z_i in table 2 are calculated for intersections of rays with the cylinder $r = r_s$, so that the curved image is presumed to lie on this cylinder. The intersection of the envelope of the family of rays with a tangent plane to $r = r_s$, normal to the mid-plane, would have less curvature.

§4. THE SOLID ANGLE OF COLLECTION

Let Ω be the maximum solid angle formed at the point source S by the beam of rays of given momentum which enters the collecting slit when the field H is correctly adjusted. A fraction $\Omega/4\pi$ of all electrons emitted with this momentum will then be collected.

We consider an idealized case in which the slit, of constant width W, is curved to fit the curvature of the image, so that the entry of rays emitted with inclination ξ_s to the mid-plane is not reduced until their sideways displacement exceeds half the length of the slit. Ω is then approximately the product of two constant angular ranges $\Delta\psi_s$ and $\Delta\xi_s$, where $\Delta\psi_s$ is the range of ψ_s, measured in the mid-plane, which enters the slit, and $\Delta\xi_s$ is the range in lateral inclination.

The relation between W and $\Delta\psi_s$ has been found graphically by drawing orbits with different ψ_s, putting $\xi_s = 0$. For each range $\Delta\psi_s$, the slit was placed across the beam in the position which made the width W a minimum. The optimum source-to-slit distance z_i varies slightly with $\Delta\psi_s$. The results for $K = 1$ and $K = \frac{1}{2}$ are given in table 3, which also gives the slit width W' of a semi-circular comparison spectrograph of the same length which collects the same range $\Delta\psi_s$.

It can be seen that, with $K = 1$, a slit of given width collects about six times the range $\Delta\psi_s$ that is collected by the comparison spectrograph with a slit of equal width.

Table 3

$K = m_t \, . \, v/H_1 e_0$	1·0				0·5	
Extreme angles of emission ψ_s (degrees)	200	190	190	180	160	140
	150	150	160	160	110	120
Range collected $\Delta\psi_s$ (degrees)	50	40	30	20	50	20
Slit width W	0·064	0·038	0·0205	0·009	0·125	0·02
Source-to-slit distance z_i	3·41	3·46	3·52	3·60	2·55	2·55
Ratio $W : z_i$	0·019	0·011	0·0058	0·0025	0·049	0·0079
$W' : z_i$ for semi-circular spectrograph	0·097	0·06	0·0341	0·0152		
Ratio $W : W$	5·0	5·5	5·8	6·1	1·91	1·93

The largest value of the inclination ξ_s which allows the ray to enter the collecting slit is set by the azimuthal angle $2\phi_0$ between the two meridional planes $\phi = \pm\phi_0$ which contain the ends of the slit. This angle, in turn, is limited by the angle of inclination of the pole-faces.

For $\xi_s = 5°$, equating ϕ_0 to the angle between the minima in the (r, ϕ) plane, we have $\phi_0 = 14°{\cdot}60$. It is clear that a higher value of ξ_s, for example $7°{\cdot}5$, could be used with steeply inclined pole-faces.

For comparison we can calculate ξ_s'. for a semi-circular spectrograph of radius ρ and slit-length l. Then $\sin\xi_s'. = l/2\pi\rho$.

Let $l = 1$ cm. and $\rho = 5$ cm. The range $\Delta\xi_s'$ is then $3°{\cdot}64$ compared with $\Delta\xi_s = 10°$.

If $\Omega = \Delta\xi_s . \Delta\psi_s$, we find that for $K = 1$ the inclined pole spectrograph has a solid angle of collection Ω about 16·5 times that of a parallel pole spectrograph with a slit of equal width and of length $\rho/5$, with the same source-image distance.

§ 5. MOMENTUM DISPERSION AND RESOLVING POWER

A knowledge of Ω makes it possible to estimate the counting rate at the peak of a β-ray line, but it does not give the shape of the line, or transmission factor, which would be a measure of the resolving power.

This shape would be difficult to compute in the present case, but we can get some interesting information from what may be called the momentum dispersion. This can be defined as the longitudinal displacement Δz_i in the position of the image I due to a fractional increase in momentum $\Delta p/p$, the field H being unchanged.

In the semi-circular spectrograph

$$z_i = 2\rho = \frac{2p}{He_0},$$

Thus

$$p\frac{dz_i}{dp} = \frac{2p}{He_0} = z_i.$$

The dispersion is proportional to the length of the spectrograph. With inclined poles, for K near to or greater than unity, ψ_s is near to π, so that

$$z_i = 2Kr_s . e^K \int_0^\pi e^{K\cos\psi} . \cos\psi . d\psi, \text{ approximately},$$

$$= K^2 r_s . e^K . \pi\left\{1 + \frac{K^2}{2.4} + \frac{K^4}{2.4^2.6} + \dots\right\}, \qquad \dots\dots(34)$$

from (27). We have $p = KH_1e_0$, giving

$$p\frac{dz_i}{dp} = \frac{Kdz_i}{dK} = K^2 r_s . e^K . \pi\left\{(K+2) + \frac{K^2}{2.4}(K+4) + \frac{K^4}{2.4^2.6}(K+6) + \text{ etc.}\dots\right\}. \qquad \dots\dots(35)$$

For $K = 1$ this gives $p\frac{dz_i}{dp} = 4{\cdot}24 z_i$, which represents 4·24 times the dispersion of a semi-circular spectrograph of the same length. Because $p\frac{dz_i}{dp}$ increases very rapidly with K, a spectrograph of high resolving power is possible.

A possible arrangement of defining baffles is shown in figure 2.

§ 6. LATERAL FOCUSING WITH BEVELLED POLES

Let the inclined plane pole-faces be modified, as in figure 5, so that the planes become parallel at a radius $r = r_p = \mathrm{OP}$ from the axis of intersection $r = 0$. The field near O is uniform and equal to H_0, whereas above P it will, after passing through a region of transition, become equal, as before, to H_1/r.

Possible advantages of such an arrangement of what may be called bevelled poles are (1) that more space is made available near the source and slit, and (2) that with a given source-to-slit length z_i, about half the flux density near the source is needed to focus a given momentum. A third advantage is that lateral focusing allows the use of a wider range $\Delta\xi_s$, so that the solid angle of collection Ω can be increased.

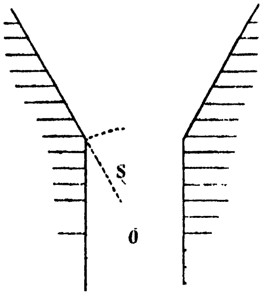


Figure 5. Profile view of the bevelled poles.

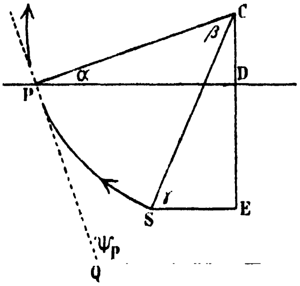


Figure 6.

Let ψ' be the angle between the axis $r = 0$ and the projection on the mid-plane of the velocity vector $\vec{v}$ at a point on the orbit. Lateral focusing is possible because, apart from small aberrations, we can find a position of the source S which allows all the members of a fan of rays with the same ψ'_s, but different inclinations ξ_s, to enter tangentially the equipotential planes $\phi = \text{constant}$ of the upper part of the field in figure 5. The tangential entry is made where the planes bend over in the transition zone near the cylinder $r = \mathrm{OP} = r_p$. The rays will remain in their respective planes until they return to $r = r_p$, when they will converge again to an image I lying in the mid-plane at $r = r_s$. Owing to small aberrations, the image will be a short line in the mid-plane rather than a point. It would be difficult to compute the orbits in the transition zone, but the uncertainty arising there should be small provided that the beam traverses the zone nearly at right angles to the axis $r = 0$.

The Lorentz force $\vec{F}$ will then be nearly parallel to that axis and will have no component in the (r, ϕ) plane, so the projections of the rays in that plane will be nearly straight when crossing the transition zone. Small variations of F with the azimuth ϕ may thus cause more disturbance of longitudinal than of lateral focusing.

§ 7. THE CONDITION FOR LATERAL FOCUSING

Let H_0 be the strength of the uniform field near the source S and let $2\phi_0$ be the angle between the inclined planes. The field at large r, $H = H_1/r$, is related to H_0 by the requirement that the pole-faces must be equipotential surfaces. This gives

$$H_1 = \frac{H_0 r_p \sin\phi_0}{\phi_0}. \qquad \ldots\ldots(36)$$

Let us make the paraxial approximation in which $\phi = \sin\phi$ and let the uniform field $H = H_0$ change abruptly at the cylinder $r = r_p$ to the decreasing field

$$H = \frac{H_1}{r} = \frac{H_0 r_p}{r}.$$

In figure 6 an orbit from S emitted in the mid-plane $\phi = 0$ crosses the boundary cylinder $r = r_p$ at P where $\psi = \psi_p$. Consider a fan of rays, emitted with different inclinations ξ_s, such that the projections on the mid-plane all have the same angle of emission ψ_s'. The projections will have radii of curvature $\rho_1' = Kr_p \cos\xi_s$ in the uniform field below r_p. Thus to the approximation that $\cos\xi_s = 1$, all the rays below P are spiral screws on a cylinder of radius $\rho_1 = Kr_p$.

Let the cylinder be straightened out so that the arc SP coincides with the straight line QP drawn along the tangent at P. Then the spiral screws coming from S become straight lines radiating from Q. If Q lies on the axis $r = 0$, the straight lines are all tangential to equipotential planes in the decreasing field above P. Thus the condition for lateral focusing is

$$\mathrm{PQ} = r_p/\sin\psi_p = \text{arc PS} = \rho_1\beta = Kr_p\beta. \qquad \ldots\ldots(37)$$

The condition cannot be exactly satisfied for large ξ_s because of the more complicated nature of the transition field and the variation of $\cos\xi_s$. Further, it is only valid for one value of ψ_s', other values giving fans of rays which, after straightening out, will radiate from points a little above or below the axis $r = 0$. This latter aberration gives rise to orbits which have small inclinations to the equipotential planes. For practical ranges of ψ_s the inclinations ξ are so small that the shortening of z_t will be negligible, and longitudinal focusing will be very little disturbed.

In order to satisfy (37) for the rays at the centre of the focused beam we must place the source S at the optimum distance above the axis $r = 0$. Let $\mathrm{DE} = C \,.\, r_p = \mathrm{CE} - \mathrm{CD} = Kr_p(\sin\gamma - \sin\alpha)$. Then

$$C = K(\sin\gamma - \sin\alpha). \qquad \ldots\ldots(38)$$

From (37)

$$\beta = \frac{\mathrm{PQ}}{Kr_p} = \frac{1}{K\sin\psi_p}.$$

We have $\alpha = \psi_p - \pi/2$, so that β can be eliminated and we can find the value of DE which gives lateral focusing for a given angle of emission $\psi_s = \gamma + \pi/2$.

This value of ψ_s should also satisfy the condition for longitudinal focusing. It seems that for $K = 1$, the fraction C should be about 0·7.

§ 8. LONGITUDINAL FOCUSING

As before, we seek for a stationary value of z_t, the distance from S at $r = r_s = r_p(1 - C)$ to the point at which the returning ray again crosses $r = r_s$. z_t will be the sum of a displacement z_p in the uniform field below $r = r_p$ and a displacement z in the decreasing field above.

$$z_p = -2(\mathrm{PD}-\mathrm{SE}) = -2\rho_0(\cos\alpha - \cos\gamma)$$
$$= -2\rho_0\left\{\cos\alpha \mp \sqrt{1-\left(\frac{C}{K}+\sin\alpha\right)^2}\right\}. \qquad \ldots\ldots(39)$$

The minus sign applies when $\gamma < \frac{\pi}{2}$.

$$z = 2Kr_p \, . \, e^{-K\cos\psi_p} \int_0^{\psi_p} e^{pK\cos\psi} \, . \cos\psi \, . \, d\psi. \qquad \ldots\ldots(40)$$

$z_i = z + z_p$ is plotted against ψ_s in figure 7. For $K = 1{\cdot}00$, $\psi_s = 160°$ gives a maximum value of z_i. Families of orbits with $K = 1{\cdot}0$ and $1{\cdot}00625$ are plotted in figure 8.

Owing to small variations of H with the azimuth ϕ in the transition zone, orbits emitted with differing inclinations ξ_s will have slightly different values of z_i. This will give rise to a lower resolving power with bevelled poles because the presence of lateral focusing will make the image length so small that it would be impossible to separate the rays with differing z_i by the use of a curved slit. It therefore seems probable that inclined plane poles have advantages where high resolving power is important, but that bevelled poles may be useful for the study of continuous spectra where they would give a higher counting rate, or for

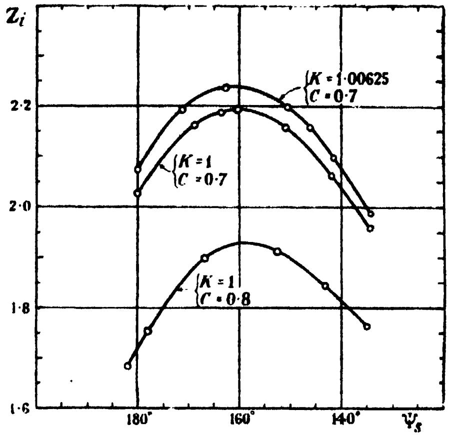
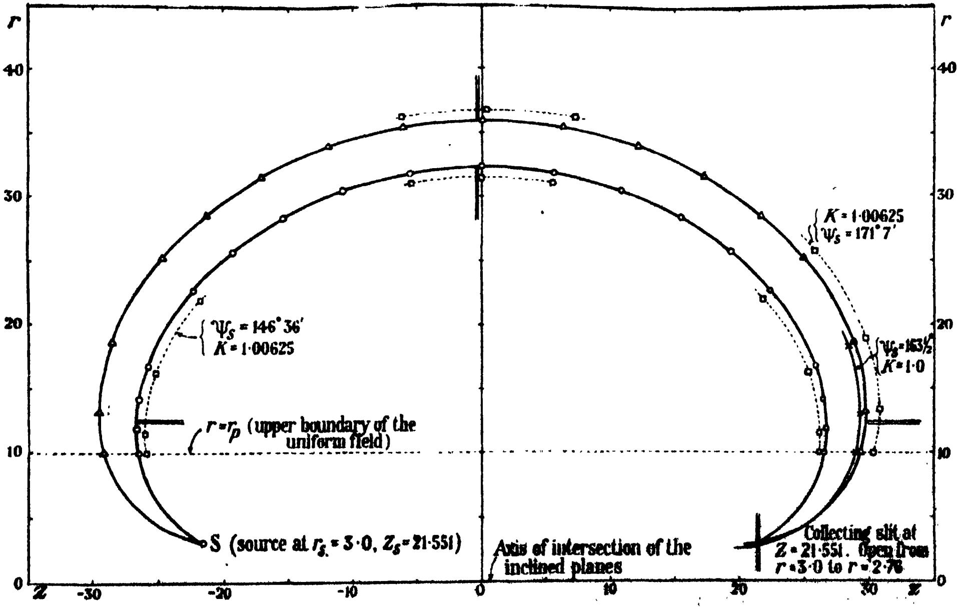


Figure 7. Longitudinal focusing between bevelled pole-faces. Source-image distance z_i against angle of emission ψ_s.



r
40
30
20
10
0
Ψ_s = 146° 36′
K = 1·00625
K = 1·00625
Ψ_s = 171° 7′
Ψ_s = 163½
K = 1·0
$r = r_p$ (upper boundary of the uniform field)
S (source at r_s = 3·0, Z_s = 21·551)
Axis of intersection of the inclined planes
Collecting slit at Z = 21·551. Open from r = 3·0 to r = 2·78
Z
-30
-20
-10
0
10
20
30
Z

Figure 8. Orbits in the mid-plane between bevelled pole-faces. The baffles are again adjusted to exclude the two ranges of rays with 0·625% higher momentum which would otherwise enter the slit. The two orbits at S have $K = 1{\cdot}0$ and $\psi_s = 150°\ 53'$ and $168°\ 40{\cdot}5'$.

high energies where the approach to saturation of the iron is a factor. The solid angle Ω could be made large enough to collect 1% of the rays emitted with a given momentum.

ACKNOWLEDGEMENT

I wish to express my thanks to Dr. J. C. P. Miller for his most valuable guidance in integration and computation.

REFERENCES

JEFFREYS, H., 1939. *Mon. Not. R. Astr. Soc.* (Geophys. Supp.), **4**, 594–615.

KORSUNSKY, M., KELMAN, V. and PETROV, B., 1945. *J. Phys. Acad. Sci. U.S.S.R.*, **9**, No. 1, 7–13.

LARMOR, J., 1884. *Proc. Lond. Math. Soc.*, **15**, 158–184.

THIBAUD, J., 1938. *Nuovo Cimento*, **15**, 313–342.

THOMSON, J. J., 1903. *Recent Advances in Electricity and Magnetism* (3rd Edition, Cambridge).

VOGES, H. and RUTHEMANN, G., 1939. *Z. Phys.*, **114**, 709–718.

WHITTAKER and ROBINSON, 1924. *The Calculus of Observations* (Blackie), p. 146.

THE PRODUCTION OF A UNIFORM MAGNETIC FIELD OVER A SPECIFIC VOLUME BY MEANS OF TWIN CONDUCTING CIRCULAR COILS

BY H. CRAIG,
Birkbeck College, London

MS. received 5 *March* 1947

ABSTRACT. An investigation is made into the uniformity of axial magnetic field attainable on the central plane between twin parallel co-axial circular coils arranged in conjunction. Information is derived and set out graphically which permits of the optimum coil separation for a given circular area with given coil diameter, as well as the highest degree of uniformity attainable over that area, being read off directly. The modification for, and consequent improved uniformity possible over, an annulus is indicated.

The investigation is then extended to volumes of cylindrical shape and some useful field characteristics, including regions of remarkably constant axial field, are considered. Two special cases are taken and axial field contours are plotted to show the deviation from uniformity of the field at any point in them. The radial field between coils with the particular separation of the two special cases is then examined.

Some important practical details are supplied. The best ratio of radial depth to axial length for the coils is determined and a method is given by which the equivalent separation of parallel coils of finite cross-section can be found very exactly by means of a search coil of particular shape.

§ 1. INTRODUCTION

PECULIARITIES associated with the magnetic field between a pair of coaxial twin coils have been noted and made use of by Rosa (1909), Llewellyn (1934), Nettleton and Balls (1935 and 1942) and Nettleton and Sugden (1939), and the Helmholtz position has not always proved the best for the requirements of these experimenters. Von Ziepel (1944) has considered the use of parallel coils for the production of a uniform field over a Wilson cloud chamber and has derived theoretically a means of establishing such a field. Two general cases are cited. Of these, the first, utilizing twin parallel coils, converges in uniformity attainable and in position upon those of the ideal Helmholtz pair as the axial length of the coil decreases. The second case, employing four axially extensive coils, is capable of producing a very homogeneous field. The calculations involved in the final solution arc indirect, and it is not always convenient to have the region subjected to the field difficult of access via the interstices of an interrupted solenoid. Tables are being published shortly by Comrie giving the values of the axial field at points off a conducting circular wire.

The object of the present communication is to show by means of these tables that a coil separation less than the Helmholtz is advantageous when the problem is that of securing optimum uniformity of field over a definite volume of specific shape, a matter of special importance in work with compound β-ray spectrometers and mass spectrographs, and to draw up data in a graphical form which will enable the appropriate separation to be read off, the deviation from uniformity to be estimated and other characteristics of the field to be appreciated.

For the production of uniform transverse or longitudinal fields in one dimension, a common requirement, as, for example, in the neutralizing of the earth's field over a single lens β-ray spectrometer, where particle paths are considerable only in one dimension, there are well-known standard methods. In the lens-magnet-lens system of compound β-ray spectroscope or in some types of mass spectrograph the particle paths are long and of appreciable extent in two dimensions. In cases such as these the effect of the earth's field can be a serious one and its neutralization over an area becomes a matter of importance. It is found convenient to compensate this effect by means of twin parallel circular co-axial coils arranged in conjunction and with an appropriate number of ampère-turns associated with them.

§ 2. NOMENCLATURE AND TABLES

Because the fields produced by all circular coils have the same shape, in the following outline all distances are given in arbitrary units of which the coil or the twin coils considered have radii of ten. Because of the axial symmetry of the field, cylindrical coordinates are used, Oz being the common axis of the coils and the origin being taken for convenience at the centre of one coil.

The values of the axial fields supplied in the tables to which reference has already been made were computed using elliptic integrals in place of the more usual zonal harmonic series and so permitting of their being established over a much wider range of r than would otherwise be possible. The field may be considered as due to a current in a circular conductor such that the ratio of ampère-turns to radius is as ten is to one, and the values of H_z are supplied in gauss to the fourth decimal place over a lattice of z and r at integral values of r and in planes of z separated by one half unit. Fields at intermediate points were calculated from these by interpolation. From the lattice of axial fields and their interpolations, other requirements can be procured as necessary. Thus, H_z at $r=r_1$ in the plane midway between two parallel circular coils in conjunction with separation Z units is twice H_z due to a single coil at $r=r_1$ when $z=Z/2$. In this plane there is no radial component of magnetic field. The H_z at any point off this plane is found by summation of the H_z due to each coil at that point.

§3. OPTIMUM COIL SEPARATION FOR A CIRCULAR AREA

Taking the axial field due to a single coil in a plane with constant z, and plotting as a percentage of H_z at $r=0$ the field difference at any particular value of r from its magnitude at $r=0$, against r, gives a smooth curve. If this is done successively for values of z from 3·8 to 5·4 a series of graphs, some of which are shown in figure 1, results. Each of the curves must, of course, intersect the ordinate at 0% when $r=0$. Curves for z planes outside the extreme ones shown are not included because they diverge very rapidly from the origin at small values of r and so are not important. From this family of curves it can be seen approximately how the uniformity of field over an area around $r=0$ varies with the z plane in which the area lies, and it becomes obvious that the best z plane for any particular area

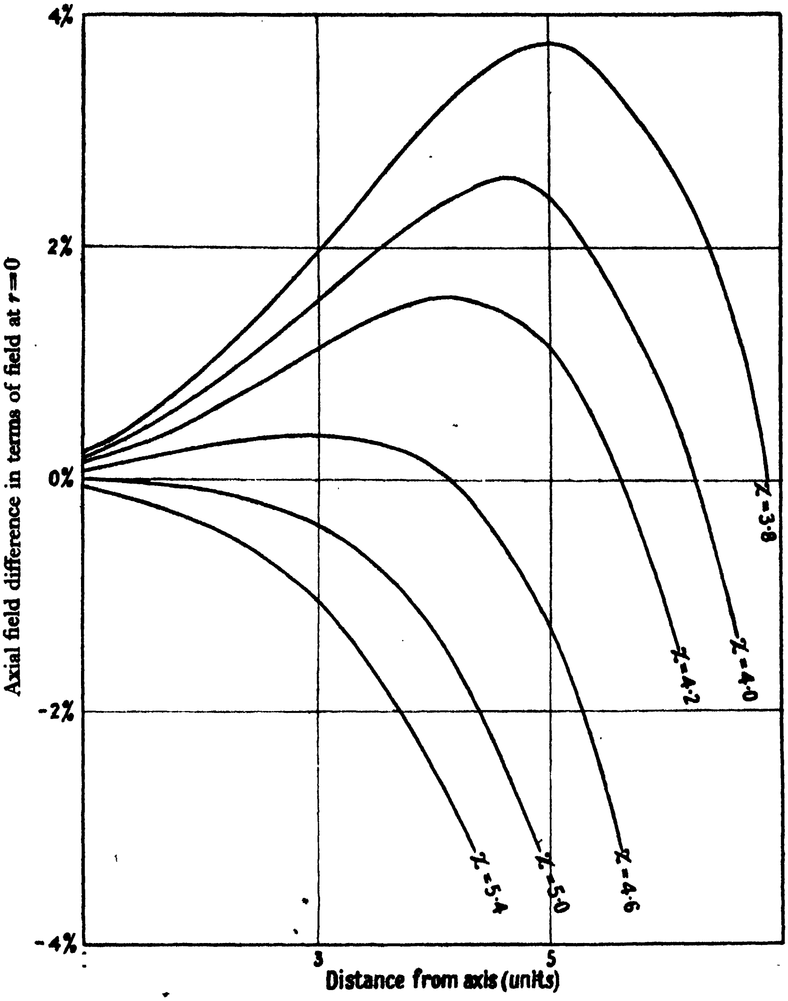


Figure 1. Variation of axial field with distance from the axis for a number of planes parallel to the coil.

depends upon the maximum value of r in that area. It is noteworthy that the curves over the range shown fall very steeply as r increases beyond certain limits, that the Helmholtz separation is only of use for small areas and that a coil separation greater than the Helmholtz is comparatively inferior.

If now the variation of H_z over the area from $r=0$ to $r=3$ be found as a percentage of the mean field over that area, and this be plotted against z for different values of z, a curve results. If this be repeated for $r_{\text{max}}=4$, $r_{\text{max}}=5$, $r_{\text{max}}=6$ and $r_{\text{max}}=7$ the family of partly coincident curves shown in figure 2 is produced and the values of z for greatest uniformity over any of the areas are then seen as minima. The twin coil separation required for any area is twice the z for the area at its minimum variation. Curves for r greater than 7 are not shown, as the field variation even at the optimum coil separation becomes too great to be of practical use. Most important features which are emphasized by these curves are that the smaller the area as compared with the coils the more critical is the distance of separation of the coils at the best position, and that the larger the area as compared with the coils the greater is the deterioration of the field on departing from the best position. Thus, an axial displacement of 0·2 units from the optimum position (brought

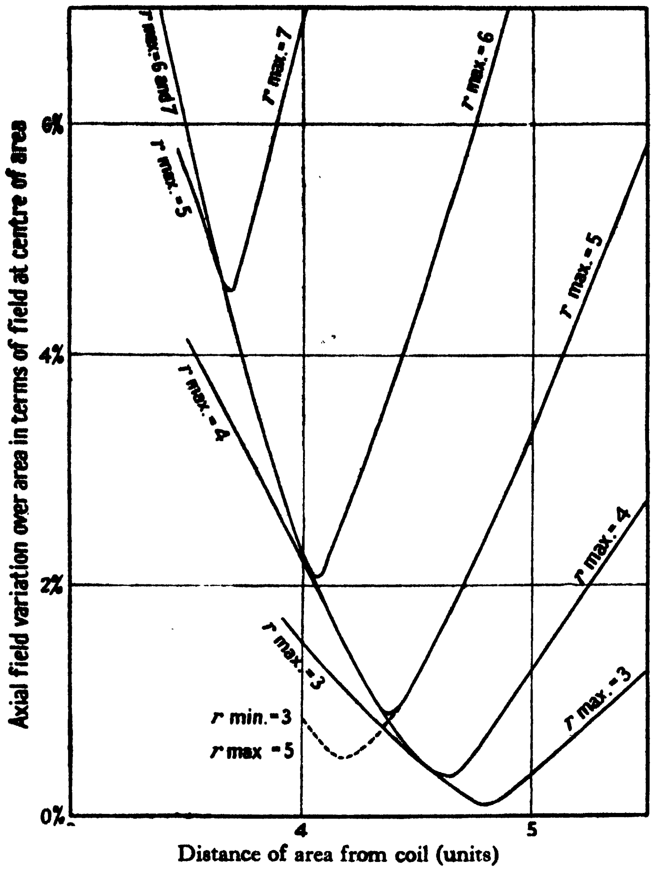


Figure 2. Variation of axial field uniformity over an area parallel to a coil against distance from coil.

about by an error of coil separation of 0·4 units) causes an increase in coarseness of the overall field of 33% for the area containing $r_{\max}=7$ and 400% for the area containing $r_{\max}=3$, whereas the actual increase in the coarseness in the former case is 1·5% as against 0·27% in the latter. Furthermore, it can be seen from figure 2 that as each curve is approximately symmetrical about its minimum, too small and too great separations of the coils are equally harmful to the field uniformity.

The z value of each minimum plotted against the r value associated with the curve provides a useful graph seen in figure 3, from which the best coil separation for any area including a known $r_{\max}$ may be procured by doubling the ordinate of the point on the graph with the known $r_{\max}$ as abscissa. Here it is seen that the Helmholtz position is approached for small areas and that the larger the area the closer must the coils be brought together. For larger areas the overall variation increases even at the best setting, as would be expected, and figure 4 shows the relation between the percentage variation of H_z over the area in the optimum z plane against $r_{\max}$ for that area. This curve shows clearly how the field uniformity attainable deteriorates as the area increases, but shows also that even between the limits $r=0$ and $r=7$, by suitable coil placing (seen to be 7·35 units apart) the field can be kept uniform to within 4·5%.

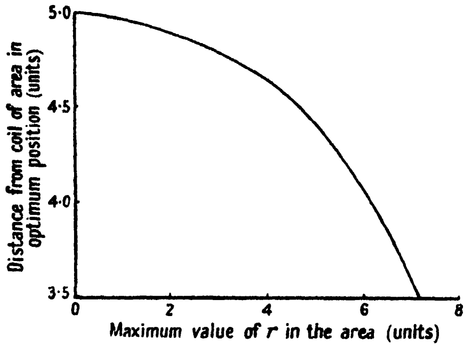


Figure 3. Optimum position for a given area.

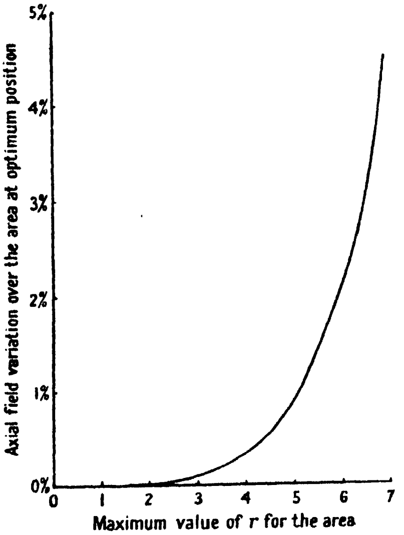


Figure 4. Field uniformity attainable over a given area.

§ 4. OPTIMUM COIL SEPARATION FOR AN ANNULUS

In the event of a region, say from $r=0$ to $r=3$, being unoccupied or of relatively little importance, as can occur for example in the particular case of a pair of long magnetic lenses at right angles in the compound β-ray spectrometer cited above, the degree of uniformity over the remainder of the region, say from $r=3$ to $r=5$, can be improved further by disregarding the variation between $r=0$ and $r=3$. In this way a more uniform field over the instruments can be obtained by suitable placing of the components over such an annulus, than by any other disposition of them. Taking this as an example and plotting the variation of axial field over the new area (i.e. between $r=3$ and $r=5$) as a percentage of the mean field over that same area, against z, and adding the resulting curve as a broken line to figure 2 gives a new minimum extending below the minimum of the $r_{\max}=5$ graph and with one arm of the former curve, if continued, almost coincident at first with the corresponding arm of the latter. It is seen that the optimum separation for the coils

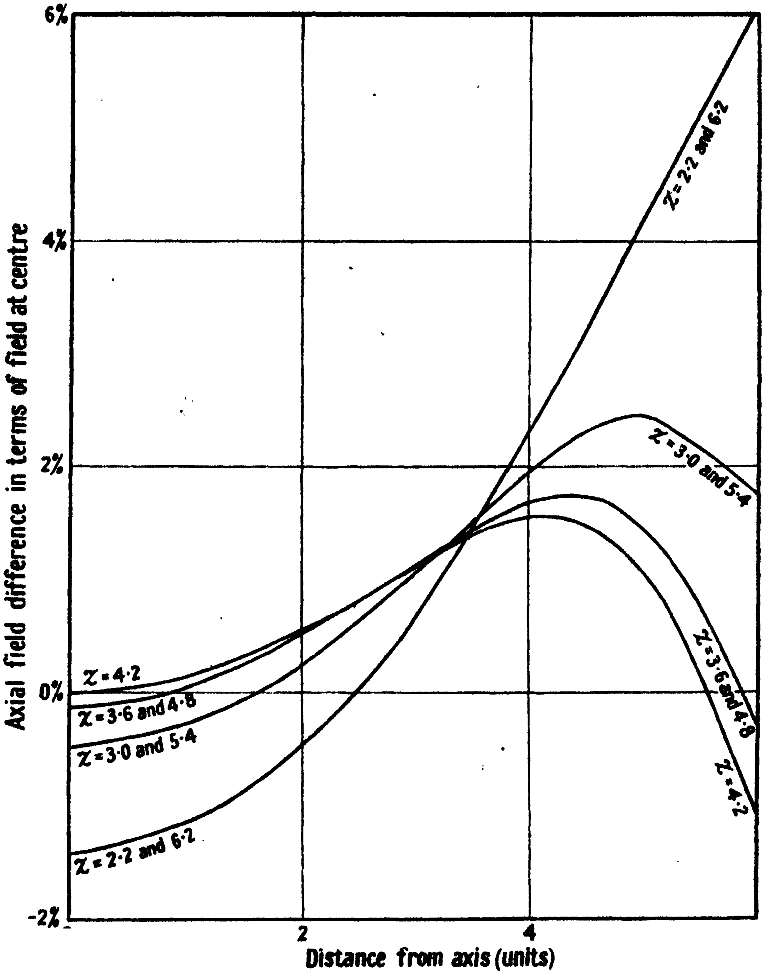


Figure 5. Variation of axial field over a number of planes between the coils.

becomes 8·38 units instead of 8·8 as would have been required for the area $r=0$ to $r=5$ and that the overall variation is now as low as 0·5%. Repeating for the areas $r=2$ to $r=5$ and $r=1$ to $r=5$ gives intermediate values both of optimum coil separation and of overall uniformity.

§ 5. FIELD IN A VOLUME

Taking H_z for positions off the central plane between the coils, it is found that the field variation from its magnitude at the same r on the central plane is small for small differences of z, but as the difference approaches two units the variation may become pronounced. Each coil separation has, of course, its own particular field form. Because of the variation of H_z with r on the central plane it is more convenient and more useful to take the H_z difference at any point in terms of the H_z on the central plane at $r=0$. A series of curves for different z planes between coils separated by 8·4 units is shown in figure 5. Other separations give families of curves of the same general character. It is observed that the curves shown within the z limits indicated, i.e. up to 2 units on either side of $z=4{\cdot}2$, intersect near $r=3{\cdot}5$ although they are not exactly concurrent. In the region of this waist the variation of H_z with z, r being constant, is a minimum and in this particular instance is of the order of 0·1 % for a range of 4 units, which is nearly one half of the distance between the coils.

The r value at the waist of the family of curves changes with the coil separation in the manner shown in figure 6. It is apparent at once that the axis of Helmholtz coils is not the best line for H_z uniformity over a range as great as two units on either side of the plane of symmetry, the best position being located near $r=1{\cdot}3$, where the uniformity is twice as good as it is along $r=0$. The axis of Helmholtz coils has a variation of 0·18 % of the centre field over the central four units of its length and an overall variation between the coils of 5·5 % of the same field. It has, of course, the advantage of being free of all radial component of magnetic field and has in addition a surrounding region extending radially as far as $r=1{\cdot}5$, in which the H_z uniformity is very good. If the region examined were reduced from that used in establishing figure 6 the curve is slightly changed. Thus for half the range, i.e. up to 1 unit from the central plane, it would intersect the ordinate at 10 units. When a z range not so close to the central plane is considered for any specific coil separation, the value of r at which the waist appears increases from its value for a z range near the central plane. Furthermore, for higher r values at the waist, i.e. for closer coils, the position is more critical and consequently the region of uniform H_z is less extensive.

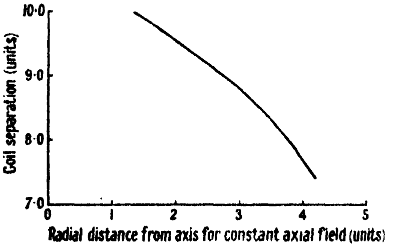


Figure 6. Variation of constant axial field position with coil separation.

For any predetermined coil separation, the way in which the axial field between them varies with z and r can best be seen by taking z and r as axes and drawing contours through points which have the same magnitude of H_z. This has been done in figure 7 (*a*) for a separation of 8·4 units over the range $r=0$ to $r=7$ and for $z=4{\cdot}2$ to $z=6{\cdot}2$. The field is symmetrical about $z=4{\cdot}2$. From the curve in figure 3 it is seen that 8·4 units is the optimum separation for the area with $r_{\max}=5{\cdot}6$ units and from the broken line addition to figure 2 for the area $r=3$ to $r=5$. For convenience the contours are drawn through the coordinates of points whose axial fields differ from that at the point $z=4{\cdot}2$, $r=0$, by the same positive or negative percentage. From such a figure as this it can be seen immediately over which region between the coils the axial field variation lies within any particular limits. It is observed that the field is very uniform on the central plane from $r=0$

to $r = 5{\cdot}6$ and that the region between $r = 3$ and $r = 5$ on $z = 4{\cdot}2$ has H_z entirely within successive contours. In the event of this latter region being the one of importance, the field would be adjusted for the mean one required over that area; contours in terms of this mean field are shown in figure 7 (*b*). They closely resemble the others in general character but have, of course, different positions, and because they are percentages of a somewhat greater field have slightly increased spacing. The particular contours for $+\frac{1}{2}\%$ and $-\frac{1}{2}\%$ have been included in this case (as broken lines) in order to indicate the trend of the field over the rather large space which otherwise would have been left ungraphed.

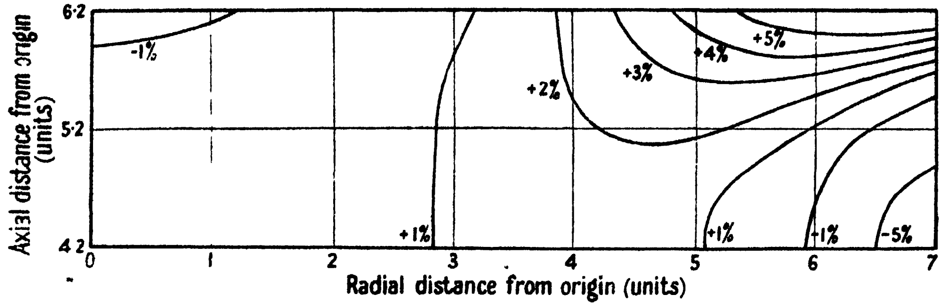


Figure 7 (*a*) Field contours in terms of field at centre between coils.

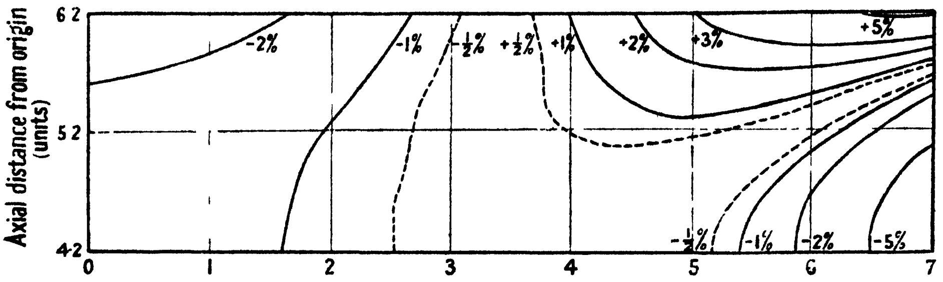


Figure 7 (*b*). Field contours in terms of mean field over annulus $r - 3$ to $r = 5$ on $z = 4{\cdot}2$.

§ 6. THE RADIAL FIELD

This brief survey of the field pertaining to the central region between parallel coils in conjunction cannot be left without some reference to the radial component of field which obviously must be present over most of the region and of which at least an approximate knowledge is advisable. Once more each coil spacing requires separate attention, as the radial field at any point is the algebraic sum of those due to each of the two coils at that point. In general, on the central plane and along the $r = 0$ axis, H_r is zero, and for small z displacements from the central plane, r being constant, the H_r at the point will be proportional to the amount of the displacement. As an individual case, the particular coil separation of 8·4 units has again been taken, and the H_r calculated for points removed from the central plane at $z = 4{\cdot}2$ by one and two units respectively, and from $r = 0$ to $r = 7$ inclusive. The calculations were made using Laplace's equation and approximate integration methods. The values of the first differential coefficients of H_z with respect to z were supplied by Comrie's tables over the same lattice of points as were the H_z

supplied. Intermediate values were obtained, as before, by interpolation. The results were calculated as percentages of H_z at $z=4{\cdot}2$, $r=0$ and are shown graphically in figure 8. The sign is arbitary as it changes everywhere in crossing the plane of symmetry. Disregarding sign, the quantities are symmetrical about $z=4{\cdot}2$. The magnitude of H_r, like the variation of H_z, is seen to be comparatively small over the important region in which H_z uniformity is being effected, and to increase rapidly as r becomes large off the central plane. Throughout the volume $r=0$ to $r=5{\cdot}6$ and $z=2{\cdot}2$ to $z=6{\cdot}2$ it is seen that H_r is never greater than the satisfactorily low proportion of 2% of the H_z at the centre.

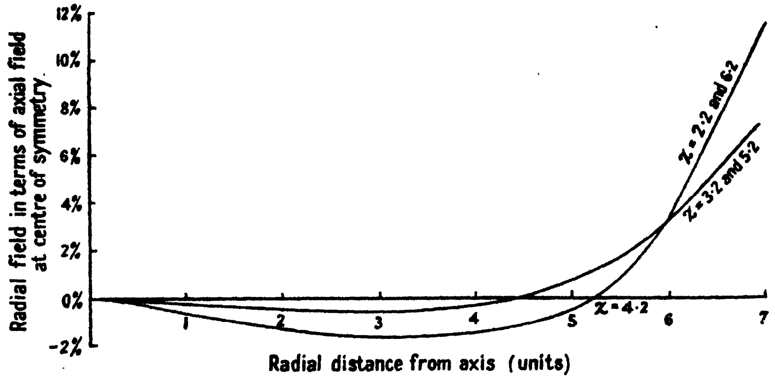


Figure 8. Variation of radial field over a number of planes between the coils.

§7. PRACTICAL CONSIDERATIONS

(a) *Coil dimensions*

All quantities from which the preceding results are derived are based on the assumption that the current is concentrated in a line, and this in practice cannot be attained perfectly. However, it can be approached very closely. The author has constructed a pair of coils 10 feet in diameter, capable on a 400-volt supply of compensating with ease the earth's field at centre between them, and they have each a cross-sectional area of less than 0·375 square inches, which is very small in comparison with the area of the section taken between the coils in a plane containing the central axis.

In such coils of mean separation $2z_1$ and mean radius a, let the radial depth and axial length be d and l respectively (see figure 9). Then considering one coil only, it follows that the optimum z plane for the area being subjected to the uniform field is distant z_1 from the plane of this coil. Taking the effect of the axial length of the coil on the axial spread of its optimum plane, it is readily seen that it is caused to be extended over a distance l

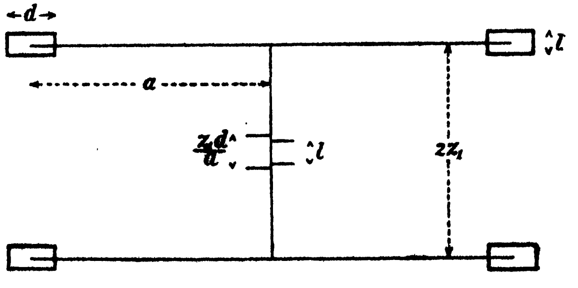


Figure 9. Coil cross-section containing central axis.

(see right-hand half of diagram). Now, taking the effect of radial depth only, the optimum plane for all turns from those of radius $a-d/2$ to those of radius $a+d/2$ must lie within limits separated by a distance $z_1 d/a$ (see left-hand half of diagram). Therefore the effect of radial depth is to the effect of axial length as coil separation is to coil diameter. It follows that for a given cross sectional area of coil (as, in order to attain maximum uniformity these limits should be equal) the axial length should be to the radial depth as the separation to be used is to the mean coil diameter. This, in addition to the comparatively small coils possible, means that the dimensions of the coils need not be such as to give rise to an appreciable effect when the field required is no greater than that of the earth.

(b) *Determination of equivalent separation*

The equivalent coil separation, Z units, can be determined experimentally by a practical method due to Dr. Nettleton and communicated privately to me by him. The fluxes G_1, G_2 produced in a search coil (see below for dimensions) placed axially at the centre of each of the twin coils in turn, by the establishment of the same current i, are determined. For identical coils they will be equal. The average flux for the two positions is $\frac{1}{2}(G_1+G_2)=ki/a$, where k is a constant and a is the coil radius. The twin coils are then arranged in opposition and the centre of the plane of symmetry between them is found by adjusting the position of the search coil, still with its axis along $r=0$, until no flux is produced on establishing a current in the twin coils. This position is critical, so can be located accurately. The coils are then arranged in conjunction and the flux G_T observed when the current i is established in the coils in series.

Then

$$G_T = \frac{2ka^2 i}{(a^2+z_1^2)^{3/2}}$$

where $2z_1$ is the coil separation. Hence

$$\frac{G_1+G_2}{G_T} = \frac{(a^2+z_1^2)^{3/2}}{a^3},$$

$$\text{i.e.}\quad \frac{z_1}{a} = \left\{\left(\frac{G_1+G_2}{G_T}\right)^{2/3} - 1\right\}^{1/2}.$$

But if $a=10$ units, then $Z=2z_1$ units and

$$Z = 20\left\{\left(\frac{G_1+G_2}{G_T}\right)^{2/3} - 1\right\}^{1/2} \text{ units.}$$

For observations such as these, in which all measurements are made at centres of symmetry of magnetic fields, the search coil should be made of a single-layer solenoid of radius R and half-length L such that $L/R=\sqrt{3}/2$, in order to eliminate the effect of its own finite dimensions. This ratio follows from the well-known fact that the magnetic field just off a centre of symmetry can be given to a first approximation by the relation

$$H(z,r) = H_0(1+bz^2-\tfrac{1}{2}br^2),$$

where H_0 is the field at the centre of symmetry, b is a constant which may be positive or negative and z and r are the coordinates of the point. The origin is taken at the centre of symmetry. If the flux in the coil is to be proportional only to H_0, then

$$\int_{-L}^{+L}\int_{0}^{R} 2\pi r(bz^2 - \tfrac{1}{2}br^2)\,dr\,dz$$

must be made zero, i.e.

$$\int_{-L}^{+L} 2\pi(\tfrac{1}{2}bz^2R^2 - \tfrac{1}{8}bR^4)\,dz = 0,$$

$$L/R = \sqrt{3}/2.$$

ACKNOWLEDGMENTS

I should like to thank Mr. R. E. Siday for his advice, for many valuable discussions and for his supply of the Comrie tables before their publication, and to thank Dr. H. R. Nettleton for his interest and helpful suggestions.

REFERENCES

Comrie, L. J. In course of publication.
Llewellyn, F. H., 1934. *Proc. Phys. Soc.*, **46**, 824.
Nettleton, H. R. and Balls, E. G., 1935. *Proc. Phys. Soc.*, **47**, 54 ; 1942. *Ibid.*, **54**, 27.
Nettleton, H. R. and Llewellyn, F. H., 1932. *Proc. Phys. Soc.*, **44**, 195.
Nettleton, H. R. and Sugden, S., 1939. *Proc. Roy. Soc.*, A, **173**, 313.
Rosa, E. B., 1909. *Bull. Bur. Standards*, **5**, 499.
von Zeipel, E., 1944. *Ark. Mat. Astr. Fys.*, **30**, A32.

DISTRIBUTION COEFFICIENTS FOR THE CALCULATION OF COLOURS ON THE C.I.E. TRICHROMATIC SYSTEM FOR TOTAL RADIATORS AT 1500–250–3500°K., AND 2360°K. (C_2=14 350)

By H. G. W. HARDING and R. B. SISSON,
National Physical Laboratory, Teddington, Middlesex

MS. received 28 *January* 1947

ABSTRACT. Tables of the distribution coefficients and the relative energy distributions are given for total radiators at 1500–250–3500° K. and 2360° K., as well as for the equal-energy stimulus and the standard illuminants A, B and C. The energy distributions required for the calculations were obtained from Planck's formula (C_2=14 350) and the distribution coefficients for the equal-energy stimulus are taken from *Condensed Tables for Colour Computation* by T. Smith (1934). By using these tables, with entries at every 0·01 μ, the labour of colour computation is only one-half of that involved by the use of

the C.I.E. standard observer tables, with entries at every 0·005 μ. In nearly every instance the calculated colour will be indistinguishable from that which would be obtained if the C.I.E. tables were used.

§ 1. INTRODUCTION

To facilitate the calculation of colours of materials from their spectral transmission or reflexion factors, it is useful to have tables which give the distribution coefficients of several illuminants. These coefficients are derived from the products of the distribution coefficients for the equal-energy stimulus and the spectral energy distributions of the illuminants. Such tables have been used at the National Physical Laboratory for many years, but failing agreement on the value that should be given to C_2 in Planck's formula they have not been published. However, there have been many requests from colorimetrists for tables, and it is considered that the value 14350 will prove the most satisfactory for this purpose. This conclusion was reached not only because 14350 was used to calculate the spectral energy distributions of the standard illuminants A, B and C in the C.I.E. tables, but also because this value had been adopted for so much work both here and in America.

When the colours of total radiators were published (Harding, 1944 and 1946) with $C_2 = 14384{\cdot}8$ (Birge, 1941), figures were given for the correction which could be applied to compensate for an alteration of C_2, and consequently the usefulness of the tables would not be affected by such an alteration. No such simple procedure is possible with tables of distribution coefficients; nevertheless, if a new value of C_2 is adopted the tables will not be without value, because they will then apply, on the new scale, to a slightly different temperature, obtained by multiplying the stated temperature by the new value of C_2 and dividing the product by 14350. For widely used illuminants, such as 2360 and 2848° K. (illuminant A) it may be preferable, in the event of an alteration of C_2, to assign a new temperature to the present distribution coefficients rather than to retain the numerical value of the temperatures at 2360 and 2848 and re-calculate the distribution coefficients.

The distribution coefficients for the equal-energy stimulus and the standard illuminants A, B and C which have been included with the newly calculated tables, have been taken from *Condensed Tables for Colour Computation* by T. Smith (1934). Offprints of these tables have not been available for some years, and it is felt that the inclusion of these widely used distribution coefficients will make table 2 more comprehensive and, therefore, more useful to colorimetrists.

§ 2. CALCULATION OF COLOURS, AND TRANSMISSION OR REFLEXION FACTORS

A colour, defined on the system recommended for use by the Commission Internationale de l'Eclairage in 1931, is expressed in the form

$$C = xX + yY + zZ,$$

where x, y and z are trichromatic coefficients whose sum is unity and where X, Y and Z denote the standard reference stimuli, X being analogous to red, Y to green and Z to blue. The values of x, y and z are obtained from the distribution coefficients for an equal-energy stimulus ($\bar{x}_\lambda$, $\bar{y}_\lambda$ and $\bar{z}_\lambda$), the spectral distribution of energy of the illuminant (E_λ) and the spectral transmission or reflexion factor of the specimen (t_λ) by making the following calculations:

$$x=\frac{1}{S}\Sigma(E_\lambda \bar{x}_\lambda t_\lambda), \qquad \ldots\ldots(1)$$

$$y=\frac{1}{S}\Sigma(E_\lambda \bar{y}_\lambda t_\lambda), \qquad \ldots\ldots(2)$$

$$z=\frac{1}{S}\Sigma(E_\lambda \bar{z}_\lambda t_\lambda), \qquad \ldots\ldots(3)$$

where $$S=\Sigma(E_\lambda \bar{x}_\lambda t_\lambda)+\Sigma(E_\lambda \bar{y}_\lambda t_\lambda)+\Sigma(E_\lambda \bar{z}_\lambda t_\lambda). \qquad \ldots\ldots(4)$$

From equations (1), (2), (3) and (4), it can be seen that, for a given illuminant, the expressions $E_\lambda \bar{x}_\lambda$, $E_\lambda \bar{y}$ and $E_\lambda \bar{z}_\lambda$ are independent of the characteristics of the material, so that if these values for any particular illuminant are tabulated, they form a set of distribution coefficients applicable to any material. As the luminosity factors of the X, Y and Z stimuli are 0, 1, 0 respectively, the transmission or reflexion factor T is given by

$$T=\Sigma(E_\lambda \bar{y}_\lambda t_\lambda)[\Sigma(E_\lambda \bar{y}_\lambda)]^{-1}. \qquad \ldots\ldots(5)$$

Hence, if the sum of the $E_\lambda \bar{y}_\lambda$ coefficients is arranged to be 100, the value $\Sigma(E_\lambda y_\lambda t_\lambda)$ represents the percentage transmission or reflexion factor without any further calculation.

At the laboratory, the computation of x, y, z and T is regularly performed in about thirty minutes with the aid of a suitable electric calculator. Recently the Hollerith section of the Mathematics Division of the Laboratory has been calculating the colours, and by this means the time taken has been approximately halved.

§ 3. PREPARATION OF THE TABLES OF DISTRIBUTION COEFFICIENTS

Tables of the distribution coefficients for total radiators at 1500–250–3500° K. and also for 2360° K. have been prepared in the following way.

The relative energy distributions of the total radiators were calculated from Planck's formula

$$E_{\lambda\theta}=C_1\lambda^{-5}[\exp.(C_2/\lambda\theta)-1]^{-1},$$

where $E_{\lambda\theta}$ is the amount of energy radiated at the absolute temperature θ between the wave-lengths $\lambda \pm d\lambda/2$ (microns); C_1 and C_2 are constants. Since only the relative spectral distribution of energy is necessary for these calculations, the value of the constant C_1 is immaterial; for convenience the energies for radiation of wave-length 0·56 μ are made equal to 100·0000. The value given to C_2 is 14350. All the energy values have been checked by forming fifth differences and are estimated to be correct to about one part in a million. These energy distributions, together with those for the standard illuminants A, B and C, are given in table 1.

The distribution coefficients for the equal-energy stimulus are not those of the C.I.E. system (Smith and Guild, 1931–2) which are quoted for the wave-lengths

0·38–0·005–0·78 μ but Smith's condensed values (T. Smith, 1934) which, quoted for wave-lengths 0·38–0·01–0·77 μ, have about half the number of entries of the C.I.E. tables.

The products of the energies given in table 1 and the appropriate distribution coefficients for the equal-energy stimulus given in table 2, columns 1, 2 and 3, were calculated and multiplied by $100[\Sigma(E_{\lambda\theta}\bar{y}_\lambda)]^{-1}$. The distribution coefficients obtained in this way for the various illuminants are tabulated in table 2.

§ 4. ACCURACY OF COLOUR CALCULATIONS USING THESE TABLES

In using these tables, for most practical purposes, the calculated colours are indistinguishable from those that would be obtained if the full C.I.E. standard tables were employed. For example, the colour of a total radiator at 2360° K. calculated by the C.I.E. tables is

$$0{\cdot}489320X + 0{\cdot}414913Y + 0{\cdot}095767Z$$

and when calculated by these tables is

$$0{\cdot}489319X + 0{\cdot}414917Y + 0{\cdot}095764Z.$$

The maximum difference in the trichromatic coefficients is 0·000004, which is much less than the minimum perceptible colour difference. Other examples, quoted by T. Smith (1934), show that for a dichroic filter which has an irregular transmission curve, the difference between the trichromatic coefficients when calculated by the condensed and the C.I.E. tables may be 0·0005; for didymium glasses with very irregular spectral curves the differences may be as much as 0·003.

§ 5. CONCLUSION

Although the use of these condensed tables in place of the full C.I.E. data may introduce significant errors in the calculated colours of some very exceptional materials, in most instances they will give results sufficiently close to the C.I.E. values for the differences to be of no visual significance.

§ 6. ACKNOWLEDGMENTS

The condensed tables of distribution coefficients for the equal-energy stimulus and for the A, B and C illuminants have been included with the approval of Mr. T. Smith, Superintendent of the Light Division of the National Physical Laboratory, and by permission of the Physical Society.

These calculations have been carried out as a part of the research programme of the National Physical Laboratory, and this paper is published by permission of the Director of the Laboratory.

REFERENCES

Birge, R. T., 1941. *Rev. Mod. Phys.*, **13**, 237.
Harding, H. G. W., 1944. *Proc. Phys. Soc.*, **56**, 305.
Harding, H. G. W., 1946. *Proc. Phys. Soc.*, **58**, 1.
Smith, T., 1934. *Proc. Phys. Soc.*, **46**, 372.
Smith, T. and Guild, J., 1931–2. *Trans. Opt. Soc.*, **33**, 73.

Table 1. Relative energy distributions

Wave-length (microns)	Temperature (° K.) ($C_2 = 14350$)			
	1500	1750	2000	2250
0·38	0·2126500	0·6756414	1·607880	3·155824
0·39	0·3561294	1·031832	2·291443	4·261929
0·40	0·5793768	1·537854	3·198011	5·651747
0·41	0·9176569	2·241006	4·377885	7·369801
0·42	1·417842	3·198327	5·886990	9·461902
0·43	2·140842	4·477287	7·786404	11·97432
0·44	3·164166	6·156287	10·14175	14·95296
0·45	4·584542	8·324897	13·02238	18·44245
0·46	6·520517	11·08388	16·50054	22·48540
0·47	9·114932	14·54488	20·65031	27·12158
0·48	12·53722	18·82992	25·54662	32·38729
0·49	16·98535	24·07058	31·26401	38·31472
0·50	22·68749	30·40693	37·87575	44·93147
0·51	29·90316	37·98631	45·45264	52·26020
0·52	38·92386	46·96187	54·06199	60·31827
0·53	50·07329	57·49097	63·76674	69·11760
0·54	63·70687	69·73346	74·62463	78·66468
0·55	80·21066	83·84996	86·68733	88·96027
0·56	100·0000	100·0000	100·0000	100·0000
0·57	123·5171	118·3403	114·6007	111·7741
0·58	151·2289	139·0231	130·5199	124·2679
0·59	183·6230	162·1941	147·7804	137·4618
0·60	221·2060	187·9916	166·3969	151·3323
0·61	264·4978	216·5442	186·3766	165·8515
0·62	314·0282	247·9703	207·7184	180·9879
0·63	370·3334	282·3764	230·4133	196·7071
0·64	433·9504	319·8566	254·4455	212·9720
0·65	505·4137	360·4913	279·7914	229·7425
0·66	585·2501	404·3470	306·4208	246·9773
0·67	673·9747	451·4756	334·2972	264·6334
0·68	772·0869	501·9142	363·3778	282·6664
0·69	880·0656	555·6851	393·6151	301·0314
0·70	998·3661	612·7957	424·9558	319·6828
0·71	1127·418	673·2385	457·3430	338·5754
0·72	1267·619	736·9920	490·7156	357·6637
0·73	1419·336	804·0205	525·0090	376·9030
0·74	1582·897	874·2745	560·1564	396·2491
0·75	1758·593	947·6918	596·0885	415·6589
0·76	1946·677	1024·198	632·7336	435·0903
0·77	2147·364	1103·708	670·0196	454·5026

Table 1. Relative energy distributions (*continued*)

Wave-length (microns)	Temperature (° K.) (C_2=14350)				
	2500	2750	3000	3250	3500
0·38	5·412436	8·415148	12·15516	16·59021	21·65663
0·39	7·001587	10·50930	14·74109	19·62634	25·08057
0·40	8·912871	12·93802	17·64894	22·95015	28·74125
0·41	11·17897	15·71932	20·88188	26·55179	32·61856
0·42	13·83054	18·86738	24·43885	30·41774	36·68986
0·43	16·89566	22·39234	28·31475	34·53127	40·93081
0·44	20·39914	26·30018	32·50061	38·87293	45·31581
0·45	24·36209	30·59258	36·98389	43·42103	49·81833
0·46	28·80156	35·26716	41·74888	48·15217	54·41175
0·47	33·73023	40·31740	46·77714	53·04164	59·06961
0·48	39·15632	45·73311	52·04785	58·06399	63·76597
0·49	45·08341	51·50054	57·53830	63·19335	68·47575
0·50	51·51062	57·60280	63·22428	68·40385	73·17510
0·51	58·43271	64·02024	69·08045	73·67004	77·84147
0·52	65·84010	70·73076	75·08079	78·96699	82·45382
0·53	73·71929	77·71032	81·19904	84·27069	86·99271
0·54	82·05318	84·93314	87·40891	89·55830	91·44037
0·55	90·82123	92·37231	93·68438	94·80816	95·78087
0·56	100·0000	100·0000	100·0000	100·0000	100·0000
0·57	109·5634	107·7878	106·3311	105·1150	104·0852
0·58	119·4833	115·7073	112·6540	110·1359	108·0254
0·59	129·7295	123·7300	118·9460	115·0469	111·8117
0·60	140·2706	131·8276	125·1859	119·8340	115·4362
0·61	151·0739	139·9727	131·3537	124·4846	118·8926
0·62	162·1062	148·1385	137·4308	128·9874	122·1763
0·63	173·3337	156·2991	143·4000	133·3332	125·2835
0·64	184·7226	164·4298	149·2457	137·5136	128·2118
0·65	196·2391	172·5071	154·9537	141·5219	130·9599
0·66	207·8501	180·5087	160·5112	145·3524	133·5276
0·67	219·5229	188·4139	165·9069	149·0010	135·9153
0·68	231·2257	196·2031	171·1310	152·4644	138·1244
0·69	242·9278	203·8583	176·1748	155·7405	140·1571
0·70	254·5995	211·3631	181·0310	158·8281	142·0159
0·71	266·2125	218·7024	185·6936	161·7270	143·7045
0·72	277·7400	225·8627	190·1577	164·4376	145·2261
0·73	289·1562	232·8317	194·4197	166·9613	146·5849
0·74	300·4372	239·5986	198·4766	169·2999	147·7856
0·75	311·5607	246·1538	202·3265	171·4559	148·8327
0·76	322·5054	252·4892	205·9684	173·4322	149·7313
0·77	333·2523	258·5979	209·4027	175·2324	150·4867

Table 1. Relative energy distributions (*continued*)

Wave-length (microns)	2360° K.	Standard illuminant		
		A	B	C
0·38	4·058004	9·79	22·40	33·00
0·39	5·371485	12·09	31·30	47·40
0·40	6·988679	14·71	41·30	63·30
0·41	8·949441	17·68	52·10	80·60
0·42	11·29332	21·00	63·20	98·10
0·43	14·05871	24·67	73·10	112·40
0·44	17·28217	28·70	80·80	121·50
0·45	20·99759	33·09	85·40	124·00
0·46	25·23566	37·82	88·30	123·10
0·47	30·02324	42·87	92·00	123·80
0·48	35·38301	48·25	95·20	123·90
0·49	41·33301	53·91	96·50	120·70
0·50	47·88649	59·86	94·20	112·10
0·51	55·05166	66·06	90·70	102·30
0·52	62·83201	72·50	89·50	96·90
0·53	71·22568	79·13	92·20	98·00
0·54	80·22638	85·95	96·90	102·10
0·55	89·82294	92·91	101·00	105·20
0·56	100·0000	100·00	102·80	105·30
0·57	110·7381	107·18	102·60	102·30
0·58	122·0143	114·44	101·00	97·80
0·59	133·8018	121·73	99·20	93·20
0·60	146·0715	129·04	98·00	89·70
0·61	158·7914	136·34	98·50	88·40
0·62	171·9273	143·62	99·70	88·10
0·63	185·4434	150·83	101·00	88·00
0·64	199·3028	157·98	102·20	87·80
0·65	213·4671	165·03	103·90	88·20
0·66	227·8972	171·96	105·00	87·90
0·67	242·5544	178·77	104·90	86·30
0·68	257·3992	185·43	103·90	84·00
0·69	272·3929	191·93	101·60	80·20
0·70	287·4969	198·26	99·10	76·30
0·71	302·6737	204·41	96·20	72·40
0·72	317·8862	210·36	92·90	68·30
0·73	333·0992	216·12	89·40	64·40
0·74	348·2781	221·66	86·90	61·50
0·75	363·3898	227·00	85·20	59·20
0·76	378·4030	232·11	84·70	58·10
0·77	393·2872	237·01	85·40	58·20

Table 2. Distribution coefficients

Equal energy stimulus			Wave-length (microns)	1500° K.		
$\bar{x}$	$\bar{y}$	$\bar{z}$		$E_{1500}\bar{x}$	$E_{1500}\bar{y}$	$E_{1500}\bar{z}$
0·0023	0·0000	0·0106	0·38	0·0000	0·0000	0·0001
0·0082	0·0002	0·0391	0·39	0·0001	0·0000	0·0005
0·0283	0·0007	0·1343	0·40	0·0006	0·0000	0·0027
0·0840	0·0023	0·4005	0·41	0·0027	0·0001	0·0128
0·2740	0·0082	1·3164	0·42	0·0135	0·0004	0·0650
0·5667	0·0232	2·7663	0·43	0·0423	0·0017	0·2063
0·6965	0·0458	3·4939	0·44	0·0768	0·0050	0·3851
0·6730	0·0761	3·5470	0·45	0·1075	0·0122	0·5664
0·5824	0·1197	3·3426	0·46	0·1323	0·0272	0·7592
0·3935	0·1824	2·5895	0·47	0·1249	0·0579	0·8222
0·1897	0·2772	1·6193	0·48	0·0828	0·1211	0·7072
0·0642	0·4162	0·9313	0·49	0·0380	0·2462	0·5510
0·0097	0·6473	0·5455	0·50	0·0077	0·5116	0·4311
0·0187	1·0077	0·3160	0·51	0·0195	1·0496	0·3292
0·1264	1·4172	0·1569	0·52	0·1714	1·9215	0·2127
0·3304	1·7243	0·0841	0·53	0·5763	3·0076	0·1467
0·5810	1·9077	0·0408	0·54	1·2893	4·2334	0·0905
0·8670	1·9906	0·0174	0·55	2·4224	5·5617	0·0486
1·1887	1·9896	0·0077	0·56	4·1406	6·9304	0·0268
1·5243	1·9041	0·0042	0·57	6·5583	8·1924	0·0181
1·8320	1·7396	0·0032	0·58	9·6506	9·1638	0·0169
2·0535	1·5144	0·0023	0·59	13·1345	9·6864	0·0147
2·1255	1·2619	0·0016	0·60	16·3776	9·7233	0·0123
2·0064	1·0066	0·0007	0·61	18·4855	9·2741	0·0064
1·7065	0·7610	0·0003	0·62	18·6667	8·3243	0·0033
1·2876	0·5311	0·0000	0·63	16·6099	6·8511	0·0000
0·8945	0·3495		0·64	13·5211	5·2830	
0·5681	0·2143		0·65	10·0015	3·7728	
0·3292	0·1218		0·66	6·7111	2·4830	
0·1755	0·0643		0·67	4·1201	1·5096	
0·0927	0·0337		0·68	2·4931	0·9063	
0·0457	0·0165		0·69	1·4010	0·5058	
0·0225	0·0081		0·70	0·7825	0·2817	
0·0117	0·0042		0·71	0·4595	0·1649	
0·0057	0·0020		0·72	0·2517	0·0883	
0·0028	0·0010		0·73	0·1384	0·0494	
0·0014	0·0006		0·74	0·0772	0·0331	
0·0006	0·0002		0·75	0·0368	0·0123	
0·0003	0·0001		0·76	0·0203	0·0068	
0·0001	0·0000		0·77	0·0075	0·0000	
21·3713	21·3714	21·3715	TOTALS	148·7536	100·0000	5·4358
$0{\cdot}33333X+0{\cdot}33333Y+0{\cdot}33334Z$			COLOUR	$0{\cdot}58521X+0{\cdot}39341Y+0{\cdot}02138Z$		

Table 2. Distribution coefficients (*continued*)

1750° K.			Wave-length (microns)	2000° K.		
$E_{1750}\bar{x}$	$E_{1750}\bar{y}$	$E_{1750}\bar{z}$		$E_{2000}\bar{x}$	$E_{2000}\bar{y}$	$E_{2000}\bar{z}$
0·0001	0·0000	0·0003	0·38	0·0002	0·0000	0·0007
0·0003	0·0000	0·0016	0·39	0·0008	0·0000	0·0038
0·0017	0·0000	0·0081	0·40	0·0038	0·0001	0·0180
0·0074	0·0002	0·0351	0·41	0·0155	0·0004	0·0737
0·0343	0·0010	0·1649	0·42	0·0678	0·0020	0·3257
0·0994	0·0041	0·4850	0·43	0·1855	0·0076	0·9053
0·1679	0·0110	0·8423	0·44	0·2969	0·0195	1·4893
0·2194	0·0248	1·1563	0·45	0·3683	0·0417	1·9414
0·2528	0·0520	1·4507	0·46	0·4039	0·0830	2·3182
0·2241	0·1039	1·4748	0·47	0·3415	0·1583	2·2475
0·1399	0·2044	1·1940	0·48	0·2037	0·2977	1·7387
0·0605	0·3923	0·8778	0·49	0·0844	0·5469	1·2238
0·0115	0·7707	0·6495	0·50	0·0154	1·0305	0·8684
0·0278	1·4989	0·4700	0·51	0·0357	1·9251	0·6037
0·2324	2·6061	0·2885	0·52	0·2872	3·2202	0·3565
0·7438	3·8818	0·1893	0·53	0·8855	4·6214	0·2254
1·5865	5·2092	0·1114	0·54	1·8223	5·9835	0·1280
2·8467	6·5359	0·0571	0·55	3·1589	7·2528	0·0634
4·6547	7·7908	0·0302	0·56	4·9962	8·3624	0·0324
7·0635	8·8234	0·0195	0·57	7·3421	9·1715	0·0202
9·9730	9·4700	0·0174	0·58	10·0500	9·5431	0·0176
13·0420	9·6181	0·0146	0·59	12·7549	9·4064	0·0143
15·6464	9·2892	0·0118	0·60	14·8652	8·8254	0·0112
17·0129	8·5353	0·0059	0·61	15·7171	7·8852	0·0055
16·5699	7·3892	0·0029	0·62	14·8986	6·6439	0·0026
14·2372	5·8725	0·0000	0·63	12·4696	5·1434	0·0000
11·2034	4·3774		0·64	9·5662	3·7377	
8·0193	3·0251		0·65	6·6807	2·5201	
5·2123	1·9285		0·66	4·2398	1·5687	
3·1026	1·1367		0·67	2·4659	0·9034	
1·8219	0·6623		0·68	1·4158	0·5147	
0·9944	0·3590		0·69	0·7560	0·2730	
0·5399	0·1944		0·70	0·4019	0·1447	
0·3084	0·1107		0·71	0·2249	0·0807	
0·1645	0·0577		0·72	0·1176	0·0412	
0·0882	0·0315		0·73	0·0618	0·0221	
0·0479	0·0205		0·74	0·0329	0·0141	
0·0223	0·0074		0·75	0·0150	0·0050	
0·0120	0·0040		0·76	0·0080	0·0026	
0·0043	0·0000		0·77	0·0028	0·0000	
136·3975	100·0000	9·5590	TOTALS	127·2603	100·0000	14·6353
$0{\cdot}55456X+0{\cdot}40658Y+0{\cdot}03886Z$			COLOUR	$0{\cdot}52610X+0{\cdot}41340Y+0{\cdot}06050Z$		

Table 2. Distribution coefficients (*continued*)

2250° K.			Wave-length (microns)	2500 °K.		
$E_{2250}\bar{x}$	$E_{2250}\bar{y}$	$E_{2250}\bar{z}$		$E_{2500}\bar{x}$	$E_{2500}\bar{y}$	$E_{2500}\bar{z}$
0·0003	0·0000	0·0015	0·38	0·0005	0·0000	0·0026
0·0015	0·0000	0·0073	0·39	0·0026	0·0001	0·0124
0·0070	0·0002	0·0334	0·40	0·0114	0·0003	0·0541
0·0272	0·0008	0·1297	0·41	0·0425	0·0012	0·2025
0·1140	0·0034	0·5474	0·42	0·1714	0·0051	0·8235
0·2982	0·0122	1·4557	0·43	0·4330	0·0177	2·1138
0·4577	0·0301	2·2959	0·44	0·6426	0·0422	3·2235
0·5454	0·0617	2·8747	0·45	0·7417	0·0839	3·9082
0·5755	0·1183	3·3029	0·46	0·7586	0·1559	4·3541
0·4690	0·2174	3·0863	0·47	0·6003	0·2782	3·9503
0·2700	0·3945	2·3047	0·48	0·3359	0·4909	2·8677
0·1081	0·7008	1·5680	0·49	0·1309	0·8486	1·8989
0·0192	1·2781	1·0771	0·50	0·0226	1·5080	1·2708
0·0429	2·3143	0·7257	0·51	0·0494	2·6631	0·8351
0·3350	3·7565	0·4159	0·52	0·3764	4·2201	0·4672
1·0035	5·2373	0·2554	0·53	1·1016	5·7490	0·2804
2·0085	6·5948	0·1411	0·54	2·1561	7·0796	0·1514
3·3894	7·7820	0·0680	0·55	3·5613	8·1766	0·0715
5·2238	8·7433	0·0338	0·56	5·3762	8·9984	0·0348
7·4872	9·3528	0·0206	0·57	7·5533	9·4353	0·0208
10·0045	9·4999	0·0175	0·58	9·8999	9·4006	0·0173
12·4047	9·1481	0·0139	0·59	12·0485	8·8854	0·0135
14·1352	8·3920	0·0106	0·60	13·4843	8·0056	0·0101
14·6233	7·3364	0·0051	0·61	13·7091	6·8778	0·0048
13·5727	6·0526	0·0024	0·62	12·5114	5·5794	0·0022
11·1304	4·5910	0·0000	0·63	10·0940	4·1635	0·0000
8·3717	3·2710		0·64	7·4731	2·9199	
5·7356	2·1636		0·65	5·0421	1·9020	
3·5729	1·3219		0·66	3·0946	1·1450	
2·0409	0·7478		0·67	1·7424	0·6384	
1·1515	0·4186		0·68	0·9694	0·3524	
0·6046	0·2183		0·69	0·5021	0·1813	
0·3161	0·1138		0·70	0·2591	0·0933	
0·1741	0·0625		0·71	0·1409	0·0506	
0·0896	0·0314		0·72	0·0716	0·0251	
0·0464	0·0166		0·73	0·0366	0·0131	
0·0244	0·0105		0·74	0·0190	0·0081	
0·0109	0·0036		0·75	0·0085	0·0028	
0·0058	0·0019		0·76	0·0044	0·0015	
0·0020	0·0000		0·77	0·0015	0·0000	
120·4007	100·0000	20·3946	TOTALS	115·1808	100·0000	26·5915
$0{\cdot}50001X+0{\cdot}41529Y+0{\cdot}08470Z$			COLOUR	$0{\cdot}47640X+0{\cdot}41361Y+0{\cdot}10999Z$		

Table 2. Distribution coefficients (*continued*)

2750° K.			Wave-length (microns)	3000° K.		
$E_{2750}\bar{x}$	$E_{2750}\bar{y}$	$E_{2750}\bar{z}$		$E_{3000}\bar{x}$	$E_{3000}\bar{y}$	$E_{3000}\bar{z}$
0·0009	0·0000	0·0041	0·38	0·0013	0·0000	0·0060
0·0040	0·0001	0·0189	0·39	0·0056	0·0001	0·0269
0·0169	0·0004	0·0801	0·40	0·0233	0·0006	0·1106
0·0608	0·0017	0·2902	0·41	0·0818	0·0022	0·3902
0·2383	0·0071	1·1447	0·42	0·3125	0·0093	1·5102
0·5849	0·0240	2·8548	0·43	0·7487	0·0307	3·6549
0·8442	0·0555	4·2350	0·44	1·0563	0·0695	5·2987
0·9489	0·1703	5·0011	0·45	1·1614	0·1313	6·1213
0·9466	0·1945	5·4330	0·46	1·1346	0·2332	6·5117
0·7312	0·3389	4·8116	0·47	0·8589	0·3981	5·6522
0·3999	0·5843	3·4131	0·48	0·4607	0·6732	3·9238
0·1524	0·9879	2·2105	0·49	0·1724	1·1174	2·5005
0·0258	1·7184	1·4482	0·50	0·0286	1·9097	1·6093
0·0552	2·9733	0·9323	0·51	0·0603	3·2483	1·0186
0·4120	4·6198	0·5115	0·52	0·4428	4·9651	0·5497
1·1833	6·1756	0·3012	0·53	1·2519	6·5333	0·3187
2·2742	7·4674	0·1597	0·54	2·3698	7·7810	0·1664
3·6910	8·4744	0·0741	0·55	3·7901	8·7020	0·0761
5·4784	9·1696	0·0355	0·56	5·5468	9·2840	0·0359
7·5722	9·4590	0·0209	0·57	7·5631	9·4476	0·0209
9·7695	9·2767	0·0171	0·58	9·6303	9·1446	0·0168
11·7099	8·6358	0·0131	0·59	11·3976	8·4054	0·0128
12·9138	7·6668	0·0097	0·60	12·4161	7·3714	0·0093
12·9433	6·4936	0·0045	0·61	12·2978	6·1698	0·0043
11·6509	5·1956	0·0020	0·62	10·9436	4·8802	0·0019
9·2752	3·8257	0·0000	0·63	8·6159	3·5538	0·0000
6·7787	2·6486		0·64	6·2294	2·4340	
4·5166	1·7038		0·65	4·1076	1·5495	
2·7387	1·0133		0·66	2·4656	0·9122	
1·5240	0·5583		0·67	1·3587	0·4978	
0·8382	0·3047		0·68	0·7402	0·2691	
0·4293	0·1550		0·69	0·3757	0·1356	
0·2192	0·0789		0·70	0·1901	0·0684	
0·1179	0·0424		0·71	0·1014	0·0364	
0·0593	0·0208		0·72	0·0506	0·0177	
0·0300	0·0107		0·73	0·0254	0·0090	
0·0154	0·0066		0·74	0·0130	0·0056	
0·0068	0·0023		0·75	0·0056	0·0019	
0·0035	0·0012		0·76	0·0029	0·0010	
0·0012	0·0000		0·77	0·0010	0·0000	
111·1625	100·0000	33·0269	TOTALS	108·0394	100·0000	39·5477

$0{\cdot}45523X+0{\cdot}40952Y+0{\cdot}13525Z$ COLOUR $0{\cdot}43637X+0{\cdot}40390Y+0{\cdot}15973Z$

Table 2. Distribution coefficients (*continued*)

3250° K.			Wave-length (microns)	3500° K.		
$E_{3250}\bar{x}$	$E_{3250}\bar{y}$	$E_{3250}\bar{z}$		$E_{3500}\bar{x}$	$E_{3500}\bar{y}$	$E_{3500}\bar{z}$
0·0018	0·0000	0·0083	0·38	0·0024	0·0000	0·0109
0·0076	0·0002	0·0361	0·39	0·0097	0·0002	0·0464
0·0306	0·0008	0·1450	0·40	0·0385	0·0010	0·1825
0·1049	0·0029	0·5002	0·41	0·1296	0·0036	0·6177
0·3921	0·0117	1·8836	0·42	0·4754	0·0142	2·2838
0·9205	0·0377	4·4936	0·43	1·0968	0·0449	5·3538
1·2736	0·0838	6·3891	0·44	1·4924	0·0981	7·4865
1·3747	0·1554	7·2451	0·45	1·5853	0·1793	8·3554
1·3192	0·2711	7·5715	0·46	1·4984	0·3080	8·5999
0·9818	0·4551	6·4612	0·47	1·0991	0·5095	7·2326
0·5182	0·7571	4·4230	0·48	0·5720	0·8358	4·8824
0·1909	1·2372	2·7685	0·49	0·2079	1·3476	3·0154
0·0312	2·0829	1·7553	0·50	0·0336	2·2397	1·8874
0·0648	3·4922	1·0951	0·51	0·0688	3·7090	1·1631
0·4695	5·2645	0·5828	0·52	0·4928	5·5253	0·6117
1·3098	6·8355	0·3334	0·53	1·3591	7·0927	0·3459
2·4477	8·0370	0·1719	0·54	2·5121	8·2483	0·1764
3·8668	8·8779	0·0776	0·55	3·9266	9·0153	0·0788
5·5918	9·3594	0·0362	0·56	5·6207	9·4077	0·0364
7·5373	9·4153	0·0208	0·57	7·5020	9·3712	0·0207
9·4915	9·0128	0·0166	0·58	9·3576	8·8857	0·0164
11·1135	8·1959	0·0124	0·59	10·8567	8·0065	0·0122
11·9818	7·1135	0·0090	0·60	11·6016	6·8878	0·0087
11·7494	5·8946	0·0041	0·61	11·2794	5·6588	0·0039
10·3546	4·6175	0·0018	0·62	9·8584	4·3963	0·0017
8·0761	3·3312	0·0000	0·63	7·6276	3·1462	0·0000
5·7864	2·2609		0·64	5·4228	2·1188	
3·7821	1·4267		0·65	3·5179	1·3270	
2·2509	0·8328		0·66	2·0785	0·7690	
1·2301	0·4507		0·67	1·1279	0·4132	
0·6649	0·2417		0·68	0·6054	0·2201	
0·3348	0·1209		0·69	0·3029	0·1094	
0·1681	0·0605		0·70	0·1511	0·0544	
0·0890	0·0320		0·71	0·0795	0·0285	
0·0441	0·0155		0·72	0·0391	0·0137	
0·0220	0·0079		0·73	0·0194	0·0069	
0·0111	0·0048		0·74	0·0098	0·0042	
0·0048	0·0016		0·75	0·0042	0·0014	
0·0024	0·0008		0·76	0·0021	0·0007	
0·0008	0·0000		0·77	0·0007	0·0000	
105·5932	100·0000	46·0422	TOTALS	103·6658	100·0000	52·4306
$0{\cdot}41963X+0{\cdot}39740Y+0{\cdot}18297Z$			COLOUR	$0{\cdot}40479X+0{\cdot}39048Y+0{\cdot}20473Z$		

Table 2. Distribution coefficients (*continued*)

2360° K.			Wave-length (microns)	Standard illuminant A		
$E_{2360}\bar{x}$	$E_{2360}\bar{y}$	$E_{2360}\bar{z}$		$E_A\bar{x}$	$E_A\bar{y}$	$E_A\bar{z}$
0·0004	0·0000	0·0019	0·38	0·0010	0·0000	0·0048
0·0020	0·0000	0·0094	0·39	0·0046	0·0001	0·0219
0·0088	0·0002	0·0419	0·40	0·0193	0·0005	0·0916
0·0335	0·0009	0·1597	0·41	0·0688	0·0019	0·3281
0·1379	0·0041	0·6627	0·42	0·2666	0·0080	1·2811
0·3551	0·0145	1·7335	0·43	0·6479	0·0265	3·1626
0·5365	0·0353	2·6914	0·44	0·9263	0·0609	4·6469
0·6299	0·0712	3·3197	0·45	1·0320	0·1167	5·4391
0·6551	0·1347	3·7598	0·46	1·0207	0·2098	5·8584
0·5266	0·2441	3·4653	0·47	0·7817	0·3624	5·1445
0·2992	0·4372	2·5538	0·48	0·4242	0·6198	3·6207
0·1183	0·7668	1·7157	0·49	0·1604	1·0398	2·3266
0·0207	1·3816	1·1643	0·50	0·0269	1·7956	1·5132
0·0459	2·4727	0·7754	0·51	0·0572	3·0849	0·9674
0·3540	3·9690	0·4394	0·52	0·4247	4·7614	0·5271
1·0489	5·4741	0·2670	0·53	1·2116	6·3230	0·3084
2·0776	6·8218	0·1459	0·54	2·3142	7·5985	0·1625
3·4711	7·9697	0·0697	0·55	3·7329	8·5707	0·0749
5·2983	8·8682	0·0343	0·56	5·5086	9·2201	0·0357
7·5237	9·3984	0·0207	0·57	7·5710	9·4574	0·0209
9·9633	9·4608	0·0174	0·58	9·7157	9·2257	0·0170
12·2468	9·0317	0·0137	0·59	11·5841	8·5430	0·0130
13·8387	8·2160	0·0104	0·60	12·7103	7·5460	0·0096
14·2008	7·1244	0·0049	0·61	12·6768	6·3599	0·0044
13·0773	5·8318	0·0023	0·62	11·3577	5·0649	0·0020
10·6429	4·3899	0·0000	0·63	8·9999	3·7122	0·0000
7·9462	3·1047		0·64	6·5487	2·5587	
5·4054	2·0390		0·65	4·3447	1·6389	
3·3440	1·2372		0·66	2·6234	0·9706	
1·8974	0·6952		0·67	1·4539	0·5327	
1·0635	0·3866		0·68	0·7966	0·2896	
0·5548	0·2003		0·69	0·4065	0·1467	
0·2883	0·1038		0·70	0·2067	0·0744	
0·1578	0·0567		0·71	0·1108	0·0398	
0·0808	0·0283		0·72	0·0556	0·0195	
0·0416	0·0148		0·73	0·0280	0·0100	
0·0218	0·0093		0·74	0·0144	0·0062	
0·0097	0·0033		0·75	0·0063	0·0021	
0·0051	0·0017		0·76	0·0032	0·0011	
0·0017	0·0000		0·77	0·0011	0·0000	
117·9314	100·0000	23·0802	TOTALS	109·8450	100·0000	35·5824

$0{\cdot}48932X+0{\cdot}41492Y+0{\cdot}09576Z$ COLOUR $0{\cdot}44757X+0{\cdot}40745Y+0{\cdot}14498Z$

Table 2. Distribution coefficients (*continued*)

Standard illuminant B			Wave-length (microns)	Standard illuminant C		
$E_B\bar{x}$	$E_B\bar{y}$	$E_B\bar{z}$		$E_C\bar{x}$	$E_C\bar{y}$	$E_C\bar{z}$
0·0025	0·0000	0·0113	0·38	0·0036	0·0000	0·0164
0·0123	0·0003	0·0585	0·39	0·0183	0·0004	0·0870
0·0558	0·0014	0·2650	0·40	0·0841	0·0021	0·3992
0·2091	0·0057	0·9970	0·41	0·3180	0·0087	1·5159
0·8274	0·0248	3·9750	0·42	1·2623	0·0378	6·0646
1·9793	0·0810	9·6617	0·43	2·9913	0·1225	14·6019
2·6889	0·1768	13·4883	0·44	3·9741	0·2613	19·9357
2·7460	0·3105	14·4729	0·45	3·9191	0·4432	20·6551
2·4571	0·5050	14·1020	0·46	3·3668	0·6920	19·3235
1·7297	0·8018	11·3825	0·47	2·2878	1·0605	15·0550
0·8629	1·2609	7·3655	0·48	1·1038	1·6129	9·4220
0·2960	1·9190	4·2939	0·49	0·3639	2·3591	5·2789
0·0437	2·9133	2·4552	0·50	0·0511	3·4077	2·8717
0·0810	4·3669	1·3694	0·51	0·0898	4·8412	1·5181
0·5405	6·0602	0·6709	0·52	0·5752	6·4491	0·7140
1·4555	7·5959	0·3705	0·53	1·5206	7·9357	0·3871
2·6899	8·8322	0·1889	0·54	2·7858	9·1470	0·1956
4·1838	9·6060	0·0840	0·55	4·2833	9·8343	0·0860
5·8385	9·7722	0·0378	0·56	5·8782	9·8387	0·0381
7·4723	9·3341	0·0206	0·57	7·3230	9·1476	0·0202
8·8406	8·3947	0·0154	0·58	8·4141	7·9897	0·0147
9·7329	7·1777	0·0109	0·59	8·9878	6·6283	0·0101
9·9523	5·9086	0·0075	0·60	8·9536	5·3157	0·0067
9·4425	4·7373	0·0033	0·61	8·3294	4·1788	0·0029
8·1290	3·6251	0·0014	0·62	7·0604	3·1485	0·0012
6·2135	2·5629	0·0000	0·63	5·3212	2·1948	0·0000
4·3678	1·7066		0·64	3·6882	1·4411	
2·8202	1·0638		0·65	2·3531	0·8876	
1·6515	0·6110		0·66	1·3589	0·5028	
0·8796	0·3223		0·67	0·7113	0·2606	
0·4602	0·1673		0·68	0·3657	0·1329	
0·2218	0·0801		0·69	0·1721	0·0621	
0·1065	0·0384		0·70	0·0806	0·0290	
0·0538	0·0193		0·71	0·0398	0·0143	
0·0253	0·0089		0·72	0·0183	0·0064	
0·0120	0·0043		0·73	0·0085	0·0030	
0·0058	0·0025		0·74	0·0040	0·0017	
0·0024	0·0008		0·75	0·0017	0·0006	
0·0012	0·0004		0·76	0·0008	0·0003	
0·0004	0·0000		0·77	0·0003	0·0000	
99·0915	100·0000	85·3094	TOTALS	98·0699	100·0000	118·2216
$0{\cdot}34842X+0{\cdot}35162Y+0{\cdot}29996Z$			COLOUR	$0{\cdot}31006X+0{\cdot}31616Y+0{\cdot}37378Z$		

COAXIAL ELECTRON LENSES

By J. W. DUNGEY* and CATHERINE R. HULL,
British Thomson-Houston Co., Rugby
* Now at Magdalene College, Cambridge

MS. received 11 *December* 1946; *read* 30 *May* 1947

ABSTRACT. The resolving power of electron microscopes is limited chiefly by the fact that the spherical aberration of electron lenses of the conventional or " coreless " type can never be eliminated. In this paper a new type of electron lenses is investigated, called *coaxial* lenses. These contain a central cylindrical conductor, surrounded by a number of annular electrodes. The electrostatic fields in such lenses can be produced by superimposing fields of a certain simple type, calculated and tabulated in this paper. This field corresponds to a " one-element " coaxial lens, containing an annular electrode in the form of a perforated disk with rounded edges, preceded and followed by cylindrical guard rings.

At least two such elements are required to correct the aberrations of an ordinary electron microscope objective to the accuracy required, and a fully satisfactory system requires a three-element correcting lens. A two-element and a three-element system are calculated in detail and their theoretical performance is discussed.

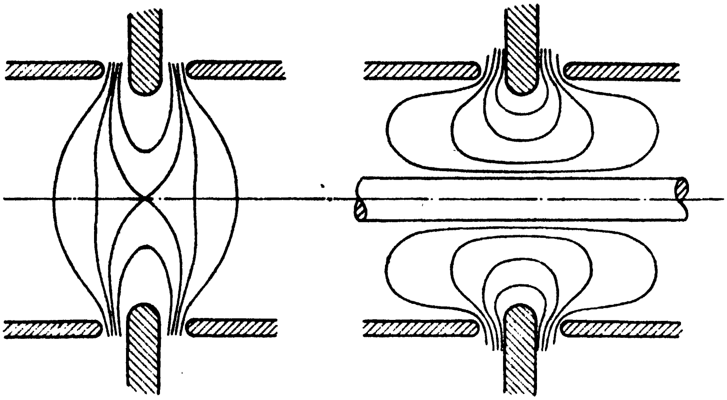
§ 1. INTRODUCTION

The resolving power of electron microscopes is at present chiefly limited by the first-order spherical aberration of the objectives, which, by a well-known theorem of Scherzer (1936), can never be eliminated in the absence of space charges. The optimum theoretical resolution has been estimated by several authors (see Zworykin *et al.*, 1945 a; Gabor, 1945) at 5 to 7 A., and recently J. Hillier (1946) has realized about 8·5 A. Further progress can be expected only from a reduction or elimination of the spherical aberration by novel means. Space-charge correction has been suggested by Gabor (1945 b) but has not yet been tried experimentally.

Another possible method is the use of coaxial electron lenses, recently suggested by Gabor (1940). Whereas in the conventional electron lenses the axis is free from electrodes (figure 1 (*a*)), coaxial lenses have one or several central electrodes or cores. Thus only an annular region is accessible to the electron beam. In reality it is not possible to utilize the full annular aperture, as the cores must be supported, but it is easy to see that a full annular aperture produces a diffraction pattern only very little smaller than a sector of this aperture with a tangential extension approximately equal to the radial width. This has the advantage that there is no need for special electron guns for the production of hollow conical beams, any ordinary electron gun with more or less circular cross-section of the beam will do, if the gun is tilted so that the beam passes obliquely through the object and eccentrically through the coaxial lens.

As the fields used in electron lenses must have accurate axial symmetry, and, moreover, as the space in electron-microscope objectives is very restricted, only the simplest arrangements are of any practical interest. The simplest type has a single straight cylindrical wire as central electrode, which must be firmly supported at both ends. The supports themselves must be in the field-free region, where the potential is the same as that of the central wire. In the following, this potential will always be assumed as zero. There may be one annular electrode with potential different from zero in the lens, or several. In the first case we speak of a one-element coaxial lens. An example of this is shown in figure 1 (*b*). According to the number of annular electrodes with non-zero potentials we speak of two-element, three-element etc. coaxial lenses.

By a zonally corrected electron lens we mean a lens in which the zonal power has a stationary value of at least the second order. If only the first differential quotient of the intercept of the rays starting from an axial object-point with respect to the initial angle of the rays is zero, that is to say, if the zonal power has a simple minimum or maximum, this means merely that an imaging effect exists in an infinitesimally narrow zone. In the case of ordinary, coreless lenses, the first-order imaging condition is always and automatically fulfilled, as the deflection angle δ is in first approximation always proportional to the off-axis distance r, but in coaxial lenses this condition is by no means automatically fulfilled. It can be satisfied only by two-element lenses or by more complex combinations.



(*a*) Coreless electron lens. (*b*) Coaxial electron lens.

Figure 1.

As in the following we shall always have to deal with objectives in which one conjugate is practically at infinity, it is convenient to characterize the zonal position by the final radius r_f at which the ray leaves the lens, practically parallel to the axis. The condition for a lens effect, that is to say, for a stationary intercept, is $d\delta/dr_f = \delta/r_f$ for the centre of the zone, $r_f = \bar{r}_f$. In a corrected combination at least the second differential coefficient must vanish, $d^2\delta/dr_f^2 = 0$, but it has been found in the course of this investigation that in order to obtain objectives better than the existing ones it is necessary to impose a third condition, $d^3\delta/dr_f^3 = 0$. It was found that these three conditions can be satisfied by certain combinations of two-element coaxial lenses with ordinary objectives, but only a three-element coaxial lens combined with a suitable coreless objective, preferably of the magnetic type, gave a convenient position of the corrected zone. This is the lens which was described by Gabor (1945). In the present paper the rather laborious steps will be described which led to this combination.

§ 2. ZONAL CORRECTION BY COAXIAL LENSES

In the following, the deflection angle δ will always be reckoned positive if the ray is deflected towards the axis, as in a condensing lens. By Scherzer's theorem,

$d^2\delta/dr_f^2$ is always positive in ordinary lenses for sufficiently small off-axial distances r_f, but there is good reason to believe that this is also true for any value of r_f, no example to the contrary being known. Thus, as the characteristic $\delta(r_f)$ of an ordinary lens is always positively curved, the first problem in a correcting lens is to produce a negative curvature. This cannot be achieved in a single lens, as shown in figure 1 (*b*), but it can be realized by a combination as shown in figure 2. Whether the trajectories are described from the left to the right or in the opposite direction is of no importance, but to simplify references it will always be assumed that the electrons move towards the right. In figure 2 they pass first through a converging lens. A single coaxial lens, whether converging or diverging, will always produce a deflection which is approximately proportional to $1/r$, where r may be the radius of the trajectory in the middle plane of the lens. Let us call the deflection produced by the first lens $\delta_1 = s_1/r_1$, where s_1 is the *strength* of the lens, a quantity of the dimension of a length. Let us now combine this with a diverging lens of strength $s_2 < 0$, so that the combination produces no deflection in a certain zone. The condition for this is

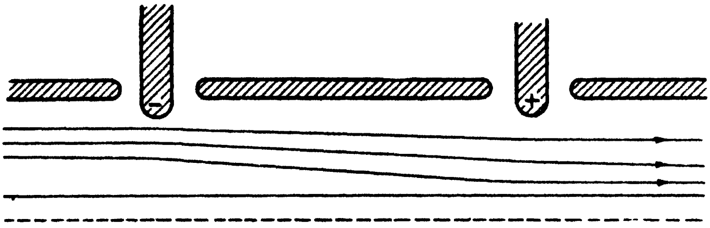


Figure 2. Correction of spherical aberration by a two-element coaxial lens.

$$s_2/s_1 = -r_2/r_1, \qquad \ldots\ldots(1)$$

but, as the first lens is convergent, $r_2 < r_1$, i.e. the divergent lens must be weaker than the convergent one.

The curvature of the characteristic contributed by each lens, $d^2\delta/dr^2$, is proportional to s/r^3, and has the same sign as the deflection. As s/r is the same for the two lenses in absolute value and r is smaller for the divergent lens, the resulting value of the curvature $d^2\delta/dr^2$ is *negative* for the combination, i.e. of a sign suitable for the correction of ordinary lenses.

This is illustrated in figure 2. It is assumed that the resulting deflection is zero for the two outer rays, hence it will be positive, i.e. towards the axis, in the case of the middle ray. The curvature had to be strongly exaggerated in order to make it visible.

The lens systems considered in the following will be similar to the one shown in figure 2 in that the resulting deflection is nearly zero. The focal power must be supplied by an ordinary lens, preferably a magnetic objective of the conventional type. Coaxial lenses which can be used by themselves as objectives are possible in principle but cannot be realized, as the fields required would cause autoelectronic discharge even under the most favourable conditions.

§ 3. THE ELECTROSTATIC FIELD IN COAXIAL ELECTRODE SYSTEMS

In the theory of ordinary coreless lenses, the potential is entirely determined if it is given along the axis. In the case of coaxial lenses, we are free to prescribe its values along two surfaces of rotational symmetry. For the above-mentioned practical reasons we need consider only the case in which the inner surface is an equipotential cylinder of radius a with a potential $\phi=0$. It is convenient to characterize the whole field by prescribing the value of ϕ as a function of the axial co-ordinate z along an outer cylinder of radius b.

In order to represent a general field, it is convenient to expand $\phi\,(b, z)$ in terms of functions which allow simple analytical representation in the annular space between the radii a and b. The conventional procedure is to expand $\phi\,(b, z)$ into a Fourier integral,

$$\phi\,(b, z)=\frac{1}{\pi}\int_0^\infty \psi\,(x)\cos zx\,dx, \qquad \ldots\ldots(2)$$

where it has been assumed for simplicity that $\phi\,(b, z)$ has the symmetry plane $z=0$ and ψ is the Fourier transform of ϕ. To obtain the solution $\phi(r, z)$ in the whole annular space between a and b we have merely to replace each coefficient $\cos zx$ by

$$\frac{I_0(ax)\,H_0(rx)-H_0(ax)\,I_0(rx)}{I_0(ax)\,H_0(bx)-H_0(ax)\,I_0(bx)}\cos zx, \qquad \ldots\ldots(3)$$

where I_0 and H_0 are the modified Bessel functions of the first and second kind of order zero. This expression satisfies Laplace's equation, and it can be seen by inspection that it satisfies the boundary conditions at $r=a$ and $r=b$. Substituting (3) into the integral (2) instead of $\cos zx$, we obtain the well-known Fourier-Bessel integral.

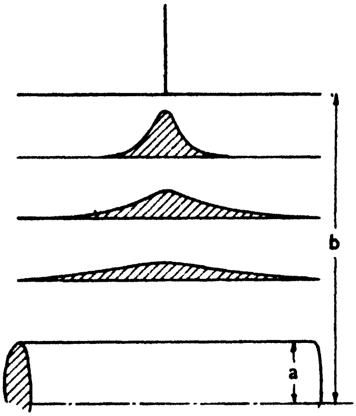


Figure 3. Potential profiles corresponding to a delta-function at the outer diameter δ.

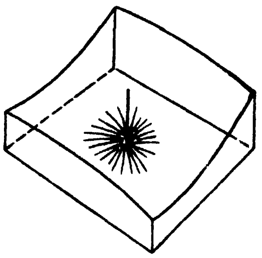


Figure 4. Illustration of relaxation method.

This procedure is not very satisfactory from a practical point of view. Each of the solutions (3) is infinitely periodic in the direction (3), and to build up functions from them which correspond to practical electrode arrangements is no easy matter. For this reason we have preferred to take as the elementary field one which corresponds to a potential of the form of a delta-function impressed on the outer electrode $r=b$. The general form of this potential distribution is illustrated in figure 3. Once this elementary solution is obtained, any prescribed potential distribution can be obtained from it by a simple integration, but in fact, not even this is necessary, as the field distribution in figure 3 is one which can be realized with good approximation with electrodes of the shape shown in figure 1 *b*, and coaxial lens systems with almost any desired property can be built up by the combination of a number of these.

Thus our first task is to calculate the potential $\phi(r, z)$ corresponding to a delta function at $r = b$, which we write, with the usual reservations, in the non-convergent form

$$\phi(b, z) = \delta(z) = \frac{1}{\pi}\int_0^\infty \cos zx\,dx. \qquad \ldots\ldots(4)$$

Replacing each $\cos zx$ by the expression (3), we obtain the corresponding $\phi(r, z)$ in the form of a Fourier-Bessel integral. This in turn can be transformed by contour integration into the following infinite series:

$$\phi(r, z) = \frac{\pi}{2}\Sigma k\,\frac{J_0(ay_k)\,N_0(ry_k) - N_0(ay_k)\,J_0(ry_k)}{\left[\dfrac{J_0(by_k)}{J_0(ay_k)} - \dfrac{J_0(ay_k)}{J_0(by_k)}\right]}\,y_k \cosh y_k z, \quad \ldots\ldots(5)$$

where the sum is to be taken over all roots y_k of the equation

$$J_0(ay_k)\,N_0(by_k) - N_0(ay_k)\,J_0(by_k) = 0. \qquad \ldots\ldots(6)$$

J_0 and N_0 are the normal Bessel functions of the first and second kinds of order zero.

The series (5), however, converges so extremely slowly that it is almost useless for practical applications. Therefore we have preferred to derive a useful approximation from equation (5), and to correct this by numerical methods.

In the series (5) we replace the cylindrical functions J_0 and N_0 by the asymptotic approximations, valid for large values of the argument

$$J_0(z) \simeq \sqrt{\frac{2}{\pi z}}\sin\left(z + \frac{\pi}{4}\right).$$

$$N_0(z) \simeq \sqrt{\frac{2}{\pi z}}\sin\left(z - \frac{\pi}{4}\right).$$

The series (5) can then be summed.

From now on we put the radius of the central wire $a = 1$, that is to say, all lengths are measured in units of a. We put $b/a = k$ and obtain the approximation

$$\phi(r, z) \simeq \frac{1}{2(k-1)}\sqrt{\frac{k}{r}}\,\frac{\sin\pi\dfrac{r-1}{k-1}}{\cosh\dfrac{\pi z}{k-1} + \cos\dfrac{r-1}{k-1}}. \qquad \ldots\ldots(7)$$

This is seen to satisfy the boundary conditions, but not the Laplace equation. It can be, however, used as the starting point of a numerical evaluation of the correct solution, by means of a relaxation method.

§ 4. NUMERICAL CALCULATION OF THE FIELD BY MEANS OF THE RELAXATION METHOD

The relaxation method has been introduced in a systematic way into a great variety of physical and engineering problems by R. V. Southwell (1940), and an application particularly useful for the present problem has been made recently by H. Motz and L. Klanfer (1946), but apart from technicalities, this method is in principle identical with J. Poincaré's *balayage* method, which starts from any function fulfilling the boundary conditions, and modifies this until Laplace's equation is satisfied, by "sweeping" the space charge gradually into the electrodes (see Picard, 1894).

The principle of the relaxation method can also be illustrated by a physical model, shown in figure 4. It is assumed that the problem is to solve Laplace's equation in two dimensions with prescribed boundary values. It is known that an evenly stretched membrane, fixed at the boundaries, assumes the shape of this function, provided that its height above the base plane varies only within narrow limits. Let us assume that in a first approximation we have arrived at provisional values of this function, and that we have pegged down the membrane to the corresponding heights at the boundaries and at certain mesh points inside the boundaries. We can now improve the solution inside a mesh, by pulling out the central peg, and letting the membrane assume its natural position. We next peg down this point to its new position, and pull out another peg. The process can be repeated any number of times, until the whole membrane is "relaxed" and the pulling out of pegs affects its shape only by a negligible amount.

In the tracing of electron trajectories to be carried out later, the values of the radial field intensity $\partial\phi/\partial r$ must be known with great accuracy. For this reason the relaxation method was applied directly not to ϕ but to $\partial\phi/\partial r$. From the results so obtained, ϕ can be determined with more than sufficient accuracy by an integration.

Differentiating Laplace's equation with respect to r we obtain

$$\frac{\partial^2}{\partial r^2}\left(\frac{\partial\phi}{\partial r}\right)+\frac{1}{r}\frac{\partial}{\partial r}\left(\frac{\partial\phi}{\partial r}\right)-\frac{1}{r^2}\left(\frac{\partial\phi}{\partial r}\right)+\frac{\partial^2}{\partial z^2}\left(\frac{\partial\phi}{\partial r}\right)=0. \qquad \ldots\ldots(8)$$

Assuming a quadratic mesh with a side length σ for each cell, and writing $\rho=r/\sigma$, i.e. measuring the radius in units of σ, we can convert equation (8) into an equation of finite differences, which expresses the value of $\partial\phi/\partial r$ for any mesh point (r, z) in terms of its values at four neighbouring mesh points. In order to make the equation clearer we write $\partial\phi/\partial r=\phi r$,

$$\phi_r(r,z)=\frac{\phi_r(r,z-\sigma)+\phi_r(r,z+\sigma)+\phi_r(r+\sigma,z)(1+1/2\rho)+\phi_r(r-\sigma,z)(1-1/2\rho)}{4+1/\rho^2}. \qquad \ldots\ldots(9)$$

A similar equation, but for ϕ instead of ϕ_r, was used by Motz and Klanfer (1946) (without the third term in equation (8)). Calculating ϕ_r instead of ϕ—which is necessary for reasons of accuracy—another difference also arises. As boundary conditions at the inner electrode we have assumed the values of ϕ_r given by the approximation (7). As these are not the exact values, what we in fact obtain by the relaxation process is a field which satisfies Laplace's equation with high accuracy, but does not correspond exactly to the initially assumed boundary conditions, that is to say, it gives no delta-function at the outer radius, but only a very sharp and narrow potential peak. This, however, is quite sufficient for all practical applications. In all the numerical calculations the outer radius was assumed as five times the inner radius. In practical applications it will be advantageous to approach the outer electrodes nearer to the axis, to save voltage. By choosing as electrode shape any of the equipotentials calculated by the relaxation method one can make sure that the potential values and field intensities calculated in this paper will apply with the indicated high accuracy.

For the start, the field values ϕ_r given by equation (7) were assumed also in the symmetry plane $z=0$, but these were slightly modified later in the relaxation process. As may be seen from figure 5, the potential field obtained finally differs but very little from that produced by a delta function at $r=5$.

Table 1 shows some of the values computed by the relaxation process and allows us to form an estimate of its accuracy. The initial values, obtained from equation (7), are underlined. Intermediate values, calculated by a first application of the relaxation method, are shown in the same row. The row below shows the values obtained finally. The mesh used in this computation was varied according to the strength of the field; in regions where it is strong, twice as fine a mesh was used as the one shown in the table. All values were computed to five figures, and the process was continued until there was no change in the fourth figure.

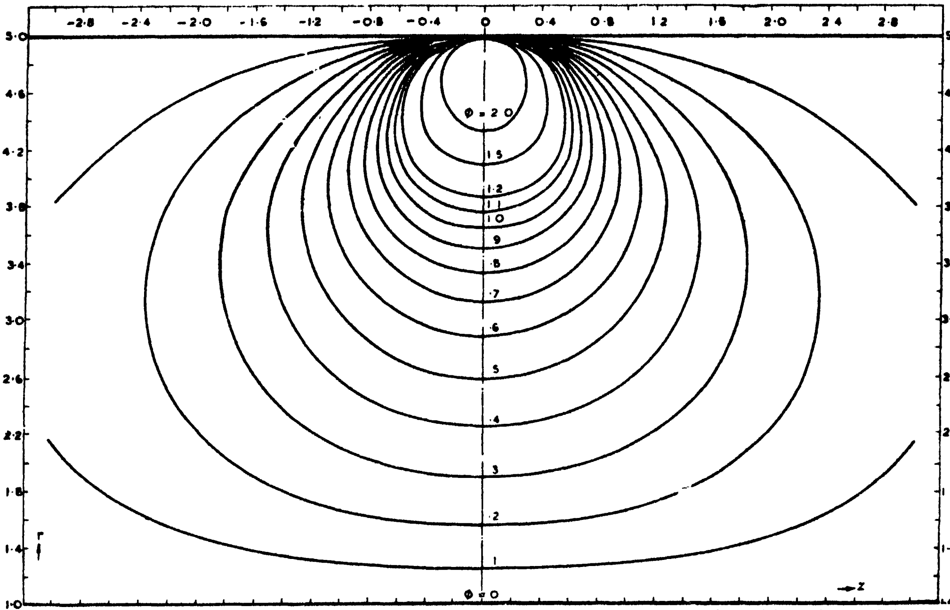


Figure 5. Potential distribution in single coaxial lens.

The accuracy of the table was checked at a number of points by using an interpolation formula more accurate than (9), by considering eight neighbours of a point Om numbered as follows:—

6
2
7 3 0 1 5
4
8

always with the spacing σ between neighbouring points. The formula is

$$\phi_r(0)=\frac{\overset{4}{\underset{1}{\Sigma}} i\,\phi_r(i)-\frac{1}{16}\overset{8}{\underset{5}{\Sigma}} i\,\phi_r(i)+\frac{3\sigma}{7r}\left[\phi_r(2)-\phi_r(4)-\frac{1}{16}(\phi_r(7)-\phi_r(8))\right]}{\frac{15}{4}+\frac{3}{4}\left(\frac{\sigma}{r}\right)^2} \qquad \ldots\ldots(9a)$$

Table 1

Computation of radial gradient by relaxation method. Comparison of initial and final values

The numbers underlined have been calculated from equation (7). The values in the top row between the calculated values were computed by the relaxation method in a first approximation. The values underneath have been calculated by the relaxation method in a second approximation, in a mesh four times finer than the one in this table.

r ↑ 3·0	·39939 ·39827	·38861 ·38665	·35758 ·35655	·31299 ·31187	·26052 ·25981	·20841 ·20858	·16007 ·16124	·11915 ·11940	·08462 ·08466	·05737 ·05763	·03700 ·03778	·02396 ·02346	·01273 ·01299	·0067 ·00622	·00090 ·00090					
2·8	·35187 ·35136	·34309 ·34314	·32176 ·32155	·28882 ·28880	·24976 ·25049	·20911 ·20729	·16992 ·17008	·13514 ·13491	·10402 ·10414	·07890 ·07876	·05832 ·05843	·04305 ·04308	·0302 ·03085	·0223 ·02228	·01549 ·01549	·01065 ·01065	·00705 ·00705	·00510 ·00510	·00274 ·00274	·00170 ·00170
2·6	·31914 ·31850	·31309 ·31311	·29732 ·29697	·27234 ·27223	·24260 ·24235	·20970 ·20973	·17745 ·17742	·14711 ·14717	·11982 ·11997	·09626 ·09633	·07606 ·07664	·06176 ·06114	·04769 ·04771	·0384 ·03769	·02938 ·02949	·02324 ·02335	·01900 ·01833	·01464 ·01450	·00132 ·01140	·00929 ·00906
2·4	·29770 ·29711	·29326 ·29369	·28066 ·28083	·26116 ·26160	·23774 ·23787	·21129 ·21112	·18419 ·18420	·15779 ·15796	·13360 ·13369	·11198 ·11189	·09261 0·9267	·07679 ·07682	·0627 ·06308	·0522 ·05211	·04263 ·04277	·03533 ·03527	·02930 ·02925	·02411 ·02407	·01987 ·01981	·01651 ·01648
2·2	·28557 ·28541	·28212 ·28222	·27236 ·27243	·25668 ·25703	·23757 ·23765	·21522 ·21505	·19215 ·19195	·16889 ·16895	·14686 ·14691	·12629 ·12635	·10735 ·10767	·09162 ·09171	·07267 ·07776	·0665 ·06603	·05549 ·05566	·04715 ·04703	·03988 ·03967	·03341 ·03348	·02830 ·02822	·02392 ·02381
2·0	·28185 ·28175	·27896 ·27901	·27074 ·27183	·25757 ·25782	·24142 ·24170	·22241 ·22258	·20196 ·20203	·1812 ·18130	·16085 ·16081	·14132 ·14114	·12272 ·12281	·10658 ·10667	·0924 ·09239	·0799 ·07986	·06828 ·06847	·05898 ·05871	·05060 ·05023	·04324 ·04298	·03695 ·03663	·03127 ·03121
1·8	·28658 ·28646	·28342 ·28410	·27711 ·27702	·26481 ·26512	·25124 ·25152	·23405 ·23458	·21526 ·21565	·19594 ·19621	·17612 ·17563	·15730 ·15598	·13845 ·13865	·12254 ·12259	1·0712 ·10755	·0941 ·09399	·08150 ·08162	·07055 ·07070	·06089 ·06089	·05252 ·05265	·04520 ·04528	·03882 ·03890
1·6	·30072 ·30075	·29864 ·29841	·29183 ·29165	·28102 ·28090	·26790 ·26799	·25186 ·25221	·23388 ·23450	·21517 ·21539	·19531 ·19414	·17648 ·17458	·15763 ·15696	·14100 ·14080	·1251 ·12473	·1097 ·10985	0·9609 ·09609	·08359 ·08366	·07255 ·07267	·06306 ·06304	·05456 ·05456	·04709 ·04710
1·4	·32692 ·32704	·32524 ·32536	·31835 ·31867	·30846 ·30804	·29440 ·29525	·27844 ·27945	·25965 ·26144	·23987 ·24113	·21876 ·22064	·20007 ·20184	·17963 ·18109	·16185 ·16231	·14304 ·14430	·12729 ·12785	·11140 ·11233	·09808 ·09838	·08558 ·08585	·07477 ·07480	·06496 ·06496	·05623 ·05626
1·2	·36981 ·37001	·36872 ·36847	·36177 ·36244	·35076 ·35281	·33593 ·33741	·31830 ·31949	·29827 ·29932	·27694 ·27777	·25490 ·25539	·23311 ·23316	·21024 ·21065	·18963 ·18991	·16878 ·16878	·14962 ·15076	·13364 ·13221	·11638 ·11610	·10155 ·10155	·08874 ·08870	·07724 ·07721	·06703 ·06698
1·0	·43905	·43635	·42839	·41552	·39842	·37775	·35436	·32914	·30290	·27638	·24925	·22498	·20098	·17855	·15782	·13888	·12157	·10635	·09262	·08046
	0	0·20	0·40	0·60	0·80	1·00	1·20	1·40	1·60	1·80	2·00	2·20	2·40	2·60	2·80	3·00	3·20	3·40	3·60	3·80

⟶ z

Table 2*a*. Radial gradient of the field of a single coaxial lens

Values of r	0	0·05	0·10	0·15	0·20	0·25	0·30	0·35	0·40	0·45	0·50	0·55	0·60	0·65	0·70	0·75	0·80	0·85	0·90	0·95	1·00
2·80	·3514	·3508	·3490	·3465	·3431	·3389	·3339	·3280	·3216	·3142	·3064	·2978	·2888	·2794	·2699	·2602	·2505	·2402	·2299	·2194	·2093
2·75	·3420	·3415	·3399	·3376	·3347	·3307	·3261	·3206	·3146	·3078	·3006	·2926	·2842	·2754	·2668	·2586	·2485	·2388	·2289	·2190	·2094
2·70	·3334	·3328	·3315	·3293	·3267	·3229	·3189	·3137	·3082	·3018	·2952	·2876	·2799	·2717	·2635	·2550	·2462	·2371	·2278	·2186	·2094
2·65	·3255	·3250	·3236	·3220	·3197	·3162	·3124	·3076	·3025	·2965	·2903	·2832	·2759	·2682	·2604	·2523	·2441	·2356	·2269	·2182	·2096
2·60	·3185	·3180	·3168	·3159	·3131	·3099	·3062	·3019	·2970	·2915	·2856	·2791	·2722	·2650	·2576	·2500	·2424	·2344	·2262	·2180	·2097
2·55	·3121	·3117	·3106	·3091	·3072	·3044	·3008	·2967	·2927	·2871	·2815	·2755	·2691	·2623	·2554	·2482	·2409	·2334	·2257	·2179	·2100
2·50	·3065	·3061	·3050	·3037	·3021	·2994	·2960	·2920	·2878	·2831	·2779	·2722	·2662	·2598	·2533	·2465	·2397	·2325	·2253	·2178	·2102
2·45	·3013	·3008	·3000	·2990	·2973	·2949	·2917	·2881	·2840	·2797	·2747	·2695	·2638	·2578	·2517	·2453	·2386	·2318	·2248	·2178	·2107
2·40	·2971	·2968	·2958	·2948	·2937	·2908	·2879	·2846	·2808	·2766	·2720	·2670	·2616	·2559	·2502	·2441	·2379	·2314	·2247	·2180	·2111
2·35	·2935	·2932	·2923	·2911	·2895	·2872	·2846	·2815	·2780	·2741	·2697	·2650	·2600	·2547	·2492	·2434	·2374	·2313	·2250	·2185	·2119
2·30	·2903	·2900	·2893	·2881	·2865	·2842	·2817	·2788	·2756	·2719	·2679	·2633	·2586	·2536	·2483	·2429	·2373	·2314	·2254	·2192	·2127
2·25	·2878	·2874	·2868	·2856	·2841	·2821	·2797	·2769	·2739	·2703	·2665	·2622	·2577	·2529	·2480	·2428	·2373	·2315	·2260	·2200	·2139
2.20	·2854	·2851	·2847	·2836	·2827	·2802	·2780	·2754	·2724	·2690	·2654	·2613	·2570	·2525	·2478	·2428	·2377	·2323	·2268	·2211	·2151
2·15	·2838	·2836	·2831	·2821	·2809	·2790	·2769	·2744	·2715	·2681	·2647	·2609	·2568	·2525	·2480	·2432	·2383	·2331	·2279	·2224	·2166
2·10	·2827	·2824	·2819	·2809	·2797	·2780	·2760	·2735	·2710	·2677	·2646	·2607	·2568	·2526	·2483	·2438	·2391	·2342	·2291	·2238	·2182
2·05	·2822	·2817	·2812	·2803	·2791	·2773	·2753	·2731	·2709	·2684	·2644	·2607	·2571	·2530	·2489	·2446	·2402	·2355	·2306	·2256	·2204
2·00	·2819	·2812	·2804	·2797	·2790	·2770	·2749	·2728	·2710	·2686	·2648	·2614	·2576	·2535	·2495	·2455	·2429	·2374	·2327	·2278	·2226
1·95	·2821	·2818	·2813	·2806	·2796	·2780	·2761	·2740	·2717	·2689	·2658	·2624	·2590	·2553	·2513	·2470	·2439	·2397	·2353	·2304	·2254
1·90	·2831	·2829	·2825	·2818	·2809	·2795	·2776	·2756	·2733	·2706	·2674	·2642	·2608	·2573	·2536	·2500	·2462	·2422	·2380	·2333	·2262
1·85	·2845	·2843	·2839	·2833	·2824	·2810	·2791	·2772	·2750	·2724	·2692	·2661	·2628	·2595	·2559	·2524	·2487	·2448	·2406	·2361	·2314
1·80	·2865	·2863	·2857	·2850	·2841	·2828	·2809	·2791	·2770	·2744	·2714	·2683	·2651	·2619	·2585	·2551	·2515	·2477	·2436	·2392	·2346
1·75	·2890	·2888	·2882	·2876	·2866	·2852	·2836	·2817	·2796	·2770	·2741	·2711	·2681	·2649	·2616	·2584	·2549	·2511	·2471	·2429	·2385
1·70	·2923	·2921	·2916	·2909	·2899	·2886	·2870	·2852	·2829	·2804	·2776	·2748	·2718	·2687	·2655	·2623	·2587	·2549	·2510	·2469	·2425
1·65	·2962	·2960	·2956	·2948	·2939	·2926	·2911	·2891	·2870	·2845	·2818	·2789	·2761	·2731	·2699	·2667	·2630	·2593	·2553	·2513	·2473
1·60	·3008	·3006	·3001	·2994	·2984	·2973	·2957	·2938	·2917	·2892	·2864	·2837	·2809	·2781	·2749	·2715	·2680	·2643	·2604	·2564	·2522
1·55	·3060	·3059	·3054	·3046	·3035	·3027	·3011	·2992	·2970	·2970	·2917	·2889	·2862	·2837	·2805	·2772	·2738	·2702	·2663	·2623	·2548
1·50	·3121	·3120	·3115	·3106	·3098	·3090	·3072	·3054	·3030	·3005	·2977	·2950	·2930	·2901	·2870	·2837	·2802	·2769	·2729	·2688	·2645
1·45	·3191	·3190	·3186	·3181	·3174	·3160	·3144	·3125	·3103	·3079	·3053	·3029	·3004	·2975	·2942	·2908	·2874	·2838	·2800	·2759	·2720
1·40	·3270	·3269	·3267	·3262	·3254	·3241	·3226	·3207	·3187	·3164	·3140	·3115	·3088	·3056	·3022	·2984	·2953	·2917	·2878	·2836	·2795
1·35	·3360	·3359	·3357	·3352	·3344	·3333	·3317	·3299	·3280	·3259	·3235	·3210	·3180	·3149	·3115	·3079	·3045	·3007	·2968	·2929	·2885
1·30	·3460	·3459	·3457	·3453	·3445	·3434	·3419	·3403	·3384	·3363	·3340	·3313	·3284	·3252	·3219	·3182	·3146	·3107	·3066	·3023	·2976
1·25	·3573	·3572	·3570	·3566	·3559	·3547	·3534	·3517	·3498	·3477	·3454	·3427	·3398	·3370	·3336	·3294	·3254	·3213	·3170	·3127	·3085
1·20	·3700	·3699	·3697	·3693	·3685	·3674	·3660	·3644	·3624	·3602	·3580	·3553	·3528	·3510	·3458	·3415	·3374	·3332	·3287	·3242	·3195
1·15	·3843	·3842	·3840	·3835	·3827	·3815	·3801	·3784	·3765	·3744	·3719	·3692	·3662	·3653	·3593	·3548	·3509	·3466	·3420	·3374	·3328
1·10	·4004	·4003	·4001	·3996	·3987	·3974	·3959	·3942	·3923	·3900	·3875	·3844	·3810	·3780	·3737	·3699	·3655	·3613	·3566	·3515	·3462
1·05	·4186	·4185	·4181	·4175	·4165	·4151	·4134	·4116	·4095	·4072	·4043	·4011	·3974	·3938	·3899	·3860	·3812	·3767	·3719	·3669	·3620
1·00	·4390	·4389	·4384	·4375	·4364	·4349	·4330	·4309	·4284	·4256	·4226	·4192	·4155	·4117	·4075	·4031	·3984	·3936	·3885	·3832	·3778

Values of Z

Table 2*b*. Radial gradient of the field of a single coaxial lens (*cont.*)

Values of *r*	1·0	1·1	1·2	1·3	1·4	1·5	1·6	1·7	1·8	1·9	2·0	2·1	2·2	2·3	2·4	2·5	2·6	2·7	2·8	2·9	3·0	3·1	3·2	3·3	3·4
2·8	·2093																								
2·7	·2094	·1914	·1739	·1573	·1414	·1264	·1124	·0997	·0879	·0773	·0679	·0598	·0523	·0454	·0395	·0345	·0301	·0260	·0225	·0196					
2·6	·2097	·1935	·1774	·1620	·1472	·1331	·1200	·1077	·0963	·0859	·0766	·0686	·0611	·0540	·0477	·0424	·0377	·0333	·0294	·0261	·0234		·0183		·0145
2·5	·2102	·1959	·1809	·1665	·1526	·1395	·1269	·1153	·1043	·0940	·0847	·0767	·0693	·0621	·0557	·0499	·0450	·0403	·0361	·0325					
2·4	·2111	·1976	·1842	·1709	·1580	·1455	·1337	·1225	·1119	·1019	·0927	·0844	·0768	·0697	·0631	·0574	·0521	·0471	·0426	·0387	·0353		·0293		·0241
2·3	·2127	·2002	·1878	·1755	·1633	·1516	·1403	·1295	·1192	·1094	·1003	·0920	·0843	·0771	·0705	·0646	·0591	·0538	·0491	·0449					
2·2	·2151	·2035	·1920	·1804	·1690	·1578	·1469	·1364	·1264	·1167	·1077	·0994	·0917	·0845	·0778	·0717	·0660	·0605	·0555	·0511	·0470		·0397		·0335
2·1	·2182	·2074	·1967	·1858	·1749	·1642	·1537	·1436	·1336	·1241	·1151	·1069	·0991	·0919	·0851	·0789	·0729	·0672	·0619	·0572					
2·0	·2226	·2124	·2020	·1916	·1813	·1710	·1608	·1508	·1411	·1317	·1228	·1145	·1067	·0993	·0924	·0860	·0799	·0739	·0683	·0633	·0587		·0502		·0430
1·9	·2262	·2182	·2083	·1984	·1883	·1782	·1681	·1583	·1486	·1394	·1306	·1224	·1145	·1069	·0999	·0932	·0868	·0806	·0748	·0696					
1·8	·2346	·2252	·2157	·2060	·1962	·1860	·1756	·1656	·1560	·1472	·1387	·1305	·1226	·1148	·1076	·1007	·0940	·0875	·0815	·0760	·0707		·0609		·0526
1·7	·2425	·2336	·2243	·2149	·2049	·1945	·1835	·1737	·1640	·1554	·1473	·1393	·1313	·1235	·1158	·1086	·1017	·0949	·0885	·0827					
1·6	·2522	·2436	·2345	·2250	·2154	·2049	·1941	·1837	·1746	·1649	·1570	·1489	·1408	·1327	·1247	·1171	·1099	·1027	·0961	·0898	·0837		·0727		·0630
1·5	·2645	·2560	·2466	·2371	·2272	·2169	·2066	·1965	·1872	·1770	·1683	·1597	·1510	·1425	·1341	·1261	·1184	·1109	·1037	·0972					
1·4	·2795	·2714	·2614	·2513	·2411	·2308	·2206	·2107	·2018	·1912	·1811	·1716	·1623	·1531	·1443	·1359	·1279	·1199	·1114	·1049	·0984		·0859		·0748
1·3	·2976	·2928	·2794	·2685	·2578	·2471	·2365	·2263	·2157	·2051	·1951	1850	·1751	·1653	·1558	·1470	·1386	·1386	·1223	·1150					
1·2	·3195	·3095	·2993	·2886	·2778	·2666	·2554	·2443	·2332	·2218	·2107	·2000	·1899	·1792	·1688	·1597	·1508	·1422	·1336	·1259	·1161		·1016		·0887
1·1	·3462	·3356	·3244	·3131	·3013	·2895	·2773	·2654	·2531	·2409	·2288	·2174	·2061	·1950	·1840	·1736	·1640	·1543	·1454	·1366					
1·0	·3778	·3661	·3544	·3418	·3291	·3160	·3029	·2896	·2764	·2628	·2493	·2371	·2250	·2130	·2010	·1898	·1786	·1682	·1582	·1484	·1389		·1216		·1063

Values of *Z*

Table 3. Values of potential in a single coaxial lens

Values of r	0	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1·0	1·1	1·2	1·3	1·4	1·5	1·6	1·7	1·8	1·9	2·0	2·1	2·2	2·3	2·4	2·5	2·6	2·7	2·8	2·9	3·0
3·0	·645	·643	·637	·628	·615	·599	·580	·559	·537	·513	·489	·464	·438	·413	·388	·364	·341	·318	·297	·276	·258	·240	·224	·208	·193	·179	·166	·154	·143	·132	·122
2·9	·606	·604	·600	·591	·580	·566	·549	·531	·511	·490	·468	·445	·422	·399	·376	·354	·332	·311	·291	·271	·253	·237	·221	·205	·191	·178	·165	·153	·142	·132	·122
2·8	·570	·569	·564	·557	·547	·535	·520	·504	·486	·467	·447	·427	·405	·384	·363	·342	·322	·302	·283	·265	·248	·232	·217	·202	·188	·176	·164	·152	·141	·131	·122
2·7	·536	·535	·531	·525	·516	·504	·491	·477	·461	·444	·426	·408	·388	·369	·349	·330	·311	·292	·275	·258	·242	·227	·212	·198	·185	·172	·161	·150	·139	·130	·120
2·6	·503	·502	·499	·493	·486	·475	·464	·451	·437	·421	·405	·388	·370	·353	·335	·317	·299	·282	·266	·249	·234	·220	·206	·193	·180	·169	·158	·147	·136	·126	·118
2·5	·472	·471	·468	·463	·456	·447	·437	·425	·413	·399	·384	·369	·352	·336	·320	·303	·287	·271	·256	·241	·226	·213	·200	·187	·175	·164	·153	·143	·133	·124	·116
2·4	·442	·441	·439	·434	·428	·420	·410	·400	·389	·376	·363	·349	·334	·319	·304	·289	·274	·259	·245	·231	·217	·205	·193	·181	·169	·159	·149	·139	·129	·121	·112
2·3	·413	·412	·410	·406	·400	·393	·384	·375	·365	·354	·342	·329	·316	·302	·288	·274	·260	·246	·233	·220	·208	·196	·185	·173	·163	·153	·143	·134	·125	·117	·109
2·2	·384	·383	·381	·378	·373	·366	·359	·350	·341	·331	·320	·309	·297	·284	·272	·259	·246	·233	·221	·209	·197	·186	·176	·165	·155	·146	·137	·128	·119	·112	·104
2·1	·356	·355	·353	·350	·346	·340	·333	·326	·317	·308	·299	·289	·277	·266	·254	·243	·231	·219	·208	·197	·186	·176	·166	·156	·147	·138	·130	·122	·114	·106	·099
2·0	·327	·327	·325	·322	·318	·313	·307	·301	·293	·285	·276	·268	·257	·247	·237	·226	·215	·204	·194	·184	·174	·165	·156	·147	·138	·130	·122	·114	·107	·100	·094
1·9	·299	·299	·297	·295	·291	·287	·281	·276	·269	·262	·254	·246	·237	·227	·218	·208	·199	·189	·180	·170	·162	·153	·145	·137	·129	·121	·141	·107	·100	·094	·087
1·8	·271	·270	·269	·267	·264	·260	·255	·250	·244	·238	·231	·224	·215	·207	·199	·190	·181	·173	·165	·156	·148	·141	·133	·126	·118	·111	·105	·098	·092	·086	·081
1·7	·242	·241	·240	·239	·236	·232	·228	·224	·219	·213	·207	·201	·193	·186	·179	·163	·156	·149	·141	·134	·127	·120	·114	·107	·101	·101	·095	·089	·084	·078	·073
1·6	·212	·212	·211	·209	·207	·204	·201	·197	·192	·187	·182	·177	·171	·164	·158	·151	·144	·138	·132	·125	·119	·113	·107	·101	·095	·090	·084	·079	·074	·070	·065
1·5	·182	·181	·181	·179	·177	·175	·172	·169	·165	·161	·156	·152	·146	·141	·136	·130	·124	·119	·114	·108	·102	·097	·092	·087	·082	·077	·073	·069	·064	·060	·056
1·4	·150	·149	·149	·148	·146	·145	·142	·139	·136	·133	·129	·126	·121	·117	·112	·108	·103	·099	·094	·089	·085	·081	·076	·072	·068	·064	·061	·057	·054	·050	·047
1·3	·116	·116	·116	·115	·114	·112	·110	·108	·106	·103	·100	·097	·094	·091	·087	·084	·080	·077	·073	·070	·066	·063	·060	·056	·053	·050	·047	·045	·042	·039	·037
1·2	·080	·080	·080	·079	·079	·078	·076	·075	·073	·071	·069	·067	·065	·063	·060	·058	·056	·053	·051	·048	·046	·044	·041	·039	·037	·035	·033	·031	·029	·027	·025
1·1	·042	·042	·042	·041	·041	·040	·040	·039	·038	·037	·036	·035	·034	·033	·032	·030	·029	·028	·026	·025	·024	·023	·022	·020	·019	·018	·017	·016	·015	·014	·013

Values of Z

in which the errors arising from the fourth-order derivatives, not considered in equation (9), have been eliminated. At about three-quarters of the points checked equation (9 *a*) gave no change, and in no case did the error exceed three units in the fourth figure. Thus table 2, which contains all values of the radial gradient required for trajectory tracing, can be considered as reliable to at least 0·1%.

The table was extended in order to compute table 3, which contains the potential values. These were obtained by integration from the values of ϕ_r and can be considered as reliable to the fifth figure. In order to draw figure 5 some values of the potential at greater radii were calculated by direct application of the relaxation method to ϕ.

Equation (7), which was used as the starting point, gave a very convenient potential scale, so that there was no need to multiply all values by any constant factor. The equipotential surface $\phi = 1$ gives a very suitable electrode shape. In the corrected objective, to be calculated later, the most convenient cathode potential was found to be -30 units. That is to say, with 60 kev. electrons, as used in many modern electron microscopes, the electrode potentials will be of the order of 1–2 kv., and the accuracy of the calculation corresponds to a few hundredths of a volt.

§ 5. NUMERICAL TRAJECTORY-TRACING

The trajectories were traced in steps of 0·2 length units (one-fifth of the wire radius). Since the angles between the trajectories and the axis never exceeded 0·02, the angles were not distinguished from their sines and the axial velocity was taken as the total velocity calculated from the potential. Errors of this kind, small as they are, cancel out almost entirely, since the main interest is in the deflections at different radii.

Let ϕ_c be the cathode potential, measured, like all other potentials, from the potential of the central wire as zero, ϕ the potential at the point of the trajectory under consideration, and β the angle of the trajectory with the axis. With the above simplifying assumptions we can write

$$\frac{d\beta}{dz} = \frac{1}{2}\frac{\partial\phi/\partial r}{(\phi - \phi_c)}, \qquad \ldots\ldots(10)$$

and if the step length s is sufficiently small, the deflection $\Delta\beta$ suffered along s is

$$\Delta\beta = \frac{s}{2}\frac{\partial\phi/\partial r}{(\phi - \phi_c)}. \qquad \ldots\ldots(11)$$

The trajectories were started parallel to the axis, and equation (11) applied to each step. After every step the new potential value was inserted, belonging to the radius $(r - s\beta)$. The values of ϕ and $\partial\phi/\partial r$ were interpolated from tables 2 and 3. Five decimals of $\partial\phi/\partial r$ were used, and three decimals for ϕ, which, added to $\phi_c = -30$, were sufficiently accurate. Seven decimals were used for the deflections. As the final deflections were of the order 0·001, and since they were the sum of a large number of steps (80 steps in two-element lens), neglecting the seventh decimal might have produced an error in the fifth decimal or third figure of these deflections. The deflections at each step were less than 0·001, so that only four figures were needed, which could be reliably obtained from the tables of $\partial\phi/\partial r$ and ϕ.

§ 6. TWO-ELEMENT LENSES

Preliminary calculations, carried out by other methods, gave a rough indication of the dimensions of a two-element correcting lens which promised zonal correction. With the cathode at -30 units, the potential impressed on the converging element was -1 unit, on the diverging element 0·86 units, and the distance between the two was 8 wire radii. Apart from the potential of the diverging element, this combination forms part of the three-element system shown in figure 6.

The results of the ray tracing are shown in the following table:—

Table 4

r_0	r_f	$\delta = \delta_1 + \delta_2$	δ_1	δ_2
2·6	2·4619	0·002371	0·015972	−0·013601
2·4	2·2556	0·002357	0·016599	−0·014242
2·2	2·0482	0·002338	0·017429	−0·015091
2·0	1·8387	0·002299	0·018520	−0·016221
1·8	1·6262	0·002233	0·019950	−0·017717

The first column gives the initial radii r_0, 4 units before the centre of the converging unit, the second gives the final radii r_f, 4 units beyond the centre of the diverging component. δ is the total deflection produced by the combination, reckoned positive if it is towards the axis. For the purpose of varying the combination, the deflection δ_1, midway between the two components, was also computed. This can be considered roughly as the deflection produced by the first (converging) lens. The last column, $\delta_2 = \delta - \delta_1$, can be similarly considered as the deflection produced by the last element, by itself.

In applying the last two columns for the calculation of new combinations, caution is needed, as there is some overlap of fields, and the deflections are not strictly additive or proportional to the lens strength. In combinations with strength ratios appreciably different from the one in the above example, $-1:0{\cdot}86$, new ray tracing may be necessary.

§ 7. FITTING A MAGNETIC LENS

In the above ray tracing, the initial rays were parallel to the axis, corresponding to a corrected image at infinity; thus the magnetic lens should be placed at the other end. The errors produced by this magnetic lens must be such that the rays leaving it tend to one object point.

The deflecting characteristic of the magnetic lens is assumed in the following form:

$$\delta_m = r\,f + C(r/f)^3. \qquad \ldots\ldots(12)$$

Hence r is the radius at which the electrons leave the magnetic lens and enter the correcting lens, that is to say, the radius which has been called r_f in table 4; f is the focal length of the magnetic lens and C is its coefficient spherical aberration. It is possible that in correcting magnetic lenses, higher terms of the series (12) are of importance, but as nothing is known of these it was thought better to leave them out of consideration.

If the magnetic lens and the coaxial lens are to form a corrected combination, the resulting deflection $\delta + \delta_m$ must be proportional to r in a certain zone, as discussed in § 1. It could not be assumed that the combination described in connection with table 4 hits this off exactly, but there were reasons, justified by the sequel, for believing that a slight variation in the strength of one of the components, say the second, would produce the desired effect. In order to measure this variation directly in potential units, we write the deflection produced by a slightly modified combination $\delta+\epsilon\delta_2/0{\cdot}86$, where ϵ is a small coefficient.

For the purpose of fitting equation (12) to the results of table 4, these results had also to be expressed in the form of a polynomial in r_f. Introducing $x=r_f-2{\cdot}0482$, i.e. measuring the radii from the central ray in table 4, we obtained by Lagrangian interpolation

$$\delta=10^{-4}(23{\cdot}38+1{\cdot}298x-2{\cdot}258x^2+1{\cdot}93x^3+1{\cdot}45x^4) \quad \ldots\ldots(13)$$

and similarly

$$-\delta_2/0{\cdot}86=10^{-4}(175{\cdot}47-54{\cdot}51x+36{\cdot}23x^2-14{\cdot}1x^3+0{\cdot}7x^4) \quad \ldots\ldots(14)$$

We want now to fit the two lenses in a zone centring on a radius $\bar{r}$, which has yet to be found. We write, therefore,

$$\delta+\delta_2/0{\cdot}86=a_0+a_1(r-\bar{r})+a_2(r-\bar{r})^2+a_3(r-\bar{r})^3+a_4(r-\bar{r})^4, \quad \ldots\ldots(15)$$

to which has to be added the deflection δ_m of the magnetic lens, given by equation (12). The resulting deflection, which may be called δ_r, must satisfy the three conditions

$$\frac{d\delta_r}{dr}=\frac{\delta_r}{r}, \qquad \frac{d^2\delta_r}{dr^2}=0, \qquad \frac{d^3\delta_r}{dr^3}=0, \quad \ldots\ldots(16)$$

which give in turn the three equations

$$2C(\bar{r}/f)^3=a_0-a_1r, \qquad 3C(\bar{r}/f)^3=-a_2\bar{r}^2, \qquad C(\bar{r}/f)^3=-a_3\bar{r}^3. \quad \ldots\ldots(17)$$

The coefficients a, however, themselves depend on $\bar{r}$. In order to give the equations in full, let us write $b_0\ldots.b_4$ for the numerical coefficients of the polynomial in equation (13) and $c_0\ldots.c_4$ for the corresponding figures in equation (14). Writing R for the radius of the middle ray from which x was measured, i.e. $x=r-R$ (R was 2·0482 in the above example), we obtain finally two equations:

$$36[b_0-b_1R+\tfrac{2}{3}b_2R^2+\epsilon(c_0-c_1R+\tfrac{2}{3}c_2R^2)](b_4+\epsilon c_4)=[3b_3R-b_2+\epsilon(c_3R-c_2)]^2, \quad \ldots\ldots(18)$$

$$R^2-r^2=\frac{3b_3R-b_2+\epsilon(3c_3R-c_2)}{6(b_5+\epsilon c_4)}. \quad \ldots\ldots(19)$$

These are the two equations for $\bar{r}$ and for ϵ. Once these two are fulfilled, the third of the equations (17) can be fulfilled by a suitable value of C/f^3, that is to say, by a certain series of magnetic lenses.

From the values of equations (13) and (14), ϵ was found to be 0·023. That is to say, the strength of the second lens had to be changed by less than 3%, which justifies *a posteriori* the method followed. The value of $\bar{r}$ was found to be 1·50. This is not a convenient value. It falls in a range in which the values of the deflection were found by extrapolation only, but even assuming that the results are reliable, the rays pass uncomfortably close to the inner wire, and the performance of the lens would be too much dependent on the precision of workmanship. For these reasons the search was continued for a three-element lens of more convenient properties.

§ 8. THREE-ELEMENT CORRECTING LENS

Rough calculations showed that a diverging element of strength 0·4 placed before the converging lens as in figure 6 should make correction possible at a radius of about two units. This element was traced on its own, with the trajectories starting parallel to the axis at 4 units from its centre, at the right. That is to say, the element was traced in opposite direction to the other two, so that it could be added on to the previous combination. A certain error arises of course from the fact that the emerging trajectories are now no longer parallel to the axis.

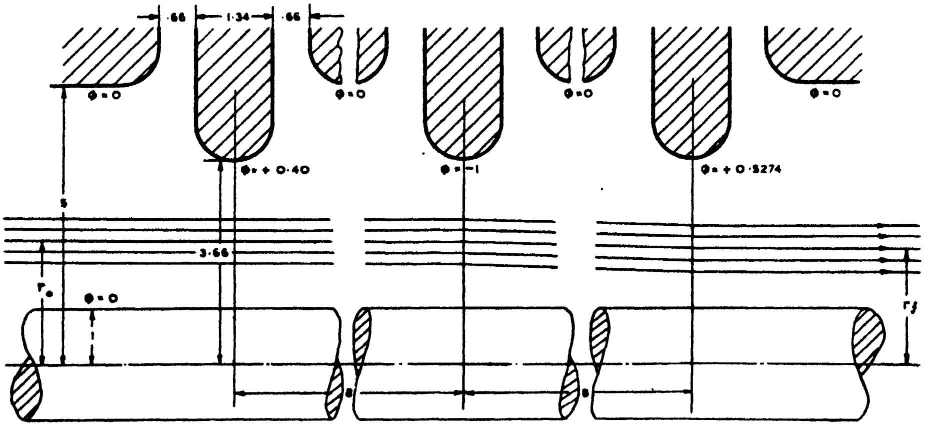


Figure 6. Electrodes and electron trajectories in three-element coaxial correcting lens. Cathode at −30 potential units.

The following values were obtained by direct ray tracing:

r_0	1·8	2·0	2·2	2·4	2·6
δ_3	−0·008068	−0·007459	−0·006972	−0·006585	−0·006274

which can be represented by the following quartic:

$$-\delta_3 = 10^{-4}(69{\cdot}72 - 20{\cdot}800x + 11{\cdot}101x^2 - 3{\cdot}35x^3 - 1{\cdot}41x^4). \qquad \ldots\ldots(20)$$

δ_3 was added to δ and the sum was used instead of δ in the calculation of ϵ and $\bar{r}$, which otherwise was carried out by the methods outlined in the last section. The result was $\epsilon = -0{\cdot}333$ and $\bar{r} = 1{\cdot}98$. The original diverging lens was thus reduced to a strength of 0·527. The whole combination with the trajectories is shown in figure 6.

The resulting characteristic of this three-element lens combination is shown in figure 7 (*a*). The magnetic lens which fits it must belong to the series

$$C/f^3 = 0{\cdot}228 \times 10^{-4}$$

or

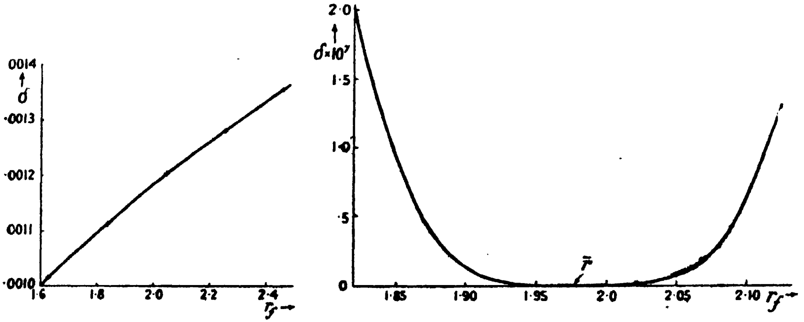
$$f = 35{\cdot}3C^{1/3} \text{ wire radii.} \qquad \ldots\ldots(21)$$

The residual error of the combination is shown in figure 7 (*b*). It has the form of a quartic parabola.

As an example let us assume that the microscope objective lens which is to be corrected is about as good as it can be made. At 60 kev. the best realizable values are about $f = 3$ mm. and $C = 0{\cdot}2$. In this case equation (21) gives a wire diameter of about 0·27 mm., which is a quite convenient value. The bore of the electrode system is about 1 mm., which means that the construction of the lens requires accuracy of a high order, but not beyond the reach of really good workshop practice.

The residual is less than 10^{-7} radian in a range of 1·98 ± 0·135 wire radii, which, with the focal length chosen, corresponds to an angular range of 0·090 ± 0·0062 radian. For a beam inside these limits the geometrical error is less than 3 A.

In order to estimate the diffraction errors, we must divide the de Broglie wavelength, which for 60 kev. electrons is about 0·05 A., by the angular range, which is about 0·012. This gives about 4 A. for the width of the first maximum in the radial diffraction figure. Two points can perhaps be separated if they are at 0·6 to 0·7 of this distance; thus we can say that the diffraction error is also 3 A. or less. Both the geometrical error and the diffractional error are less than can ever be realized in uncorrected electron microscopes of the current type.



(*a*) Characteristic of electrostatic correcting lens.

(*b*) Residual error of combination with magnetic lens.

Figure 7.

It is impossible to foresee at present how much of the calculated improvement will be realizable in practice. Evidently workmanship of the highest order will be required, and very careful focusing, much more careful than in microscopes of the current type, as the focal depth of the zonally corrected objective will be some 30 to 50 times smaller. On the other hand, this small focal depth might open up the new possibility of exploring objects in depth with an accuracy of the order of 100 A.

ACKNOWLEDGMENTS

These calculations were carried out at the suggestion and under the constant supervision of Dr. D. Gabor.

The authors wish to express their thanks to the Directors of the British Thomson-Houston Co. for permission to publish this paper.

REFERENCES

Gabor, D., 1940. *Nature, Lond.*, **158,** 198; B.T.H. Co. Brit. Patent Appln. No. 22053/45.

Gabor, D., 1945 a. *The Electron Microscope* (London : Hulton Press), Chapter 7.

Gabor, D., 1945 b. *Proc. Roy. Soc.*, A, **183,** 436.

Hillier, J., 1946. *J. Appl. Phys.*, **17,** 307, 411.

Motz, H. and Klanfer, Laura, 1946. *Proc. Phys. Soc.*, **58,** 30.

Picard, E., 1894. *Traité d'Analyse* (Paris : Gauthier Villars), Vol. II.

Scherzer, O., 1936. *Z. Phys.*, **101,** 593.

Southwell, R. V., 1940. *Relaxation Methods in Engineering Science* (Oxford : The University Press).

Zworykin, Morton, Ramberg, Hillier and Vance, 1945. *Electron Optics and the Electron Microscope* (New York : John Wiley), Chapter 19.

NOTE ON APLANATIC LENSES FOR UNIT MAGNIFICATION

By T. SMITH, F.R.S.,
National Physical Laboratory, Teddington

MS. received 30 *December* 1946

ABSTRACT. A formal solution is given for the construction of thin symmetrical cement triple objectives corrected for unit magnification. To any symmetrical objective with a diverging lens between two converging lenses there corresponds another with a converging lens between two diverging lenses, and, with the same kind of glass for the exterior lenses, the curvatures and glass properties for the central lens of one objective can be written down when those of the other objective are known. Both objectives are free from coma and have the same amounts of chromatic and spherical aberrations.

§ 1. INTRODUCTION

In lenses to be used at unit magnification coma is usually eliminated by using a symmetrical form of construction. A widely-used form consists of an equi-convex lens of a crown glass enclosed by two similar meniscus flint lenses, the three being cemented together. An alternative form, which has been used less extensively, is constructed by cementing an equi-concave flint lens to two similar converging crown lenses. A pair of glasses suitable for securing freedom from chromatic as well as spherical aberration in one type of construction has not the properties required for lenses of the other kind, and the design of one type appears to be distinct from that of the other type. But on looking at the results of some numerical calculations I noticed a connexion which could not be accidental. If a combination of each kind of the same focal length is made with the external lenses of the same glass, the curvatures of the external surfaces of both forms will have the same value if ω_2, ω_1, ω_2' are in arithmetical progression, where ω_1 is the reciprocal of the refractive index of the glass used externally, and ω_2 and ω_2' are the reciprocals of the indices of the glasses used for the central lenses in the two forms. This coincidence implies that a simple relation subsists between the internal curvatures of the two forms, and also a relation between the dispersive properties of the three glasses. But the observation is chiefly of interest as it suggests that lenses of these types can be computed with exceptional facility.

§2. GENERAL THEORY

To investigate this the notation and formulae employed in a recent paper (Smith, 1945) are adopted, so that repetition is unnecessary. Taking unity as the power of the complete lens, and j as the sum of the power and the total curvature, we have

$$\alpha = -j_1 j_2(\omega_1-\omega_2) = (j-j_2)j_2(\omega_2-\omega_1),$$

and

$$\gamma+\gamma_1 = j^2+\alpha(j+j_2)\omega_1 = j^2+(j^2-j_2^2)\omega_1 j_2(\omega_2-\omega_1). \qquad \ldots\ldots(1)$$

But since $j=j_1+i_2$, and the power of the lens is $j_1(1-\omega_1)+j_2(1-\omega_2)$,

$$j_1=\frac{j(1-\omega_2)-1}{\omega_1-\omega_2},\quad j_2=\frac{j(1-\omega_1)-1}{\omega_2-\omega_1},\qquad \ldots\ldots(2)$$

so that, when j and ω_1 are given, the value of $\gamma+\gamma_1$ is independent of the sign of j_2, i.e. of $\omega_1-\omega_2$. This verifies the numerical result observed, for the condition to be satisfied is $\gamma+\gamma_1=0$. It will be noted that the equality of curvature holds for equal amounts of spherical aberration when the ω's form an arithmetical progression. As the curvatures of the surfaces are

$$\tfrac{1}{2}(j-1),\quad \tfrac{1}{2}j_2\omega_2,\quad -\tfrac{1}{2}j_2\omega_2,\quad -\tfrac{1}{2}(j-1),$$

and j_2 is merely reversed in sign when ω_2 is replaced by ω_2', the connexion between the curvatures of the cemented surfaces is simple. Also since the range of values of ω with the glasses normally employed in optical instruments is small, the differences between the curvatures of these surfaces in the two forms will usually be small. So far as manufacture is concerned, one type of lens has no special advantage over the other. The tendency is for the shallower curves to be obtained when $\omega_1>\omega_2$, i.e. when a flint glass is cemented between two crown lenses. In practice the preference for one type rather than another will depend on the kinds of glass available and the higher-order aberrations, which are not considered here.

The dispersive properties of a glass may be represented by the value of $\omega/\delta\omega$, which may be denoted by u. The connexion between this and the quantity ν tabulated in glass lists is $\nu=(1-\omega)u$. Formally there is a change of sign, but this may be neglected as the sign is purely a matter of convention. The condition for freedom from chromatic aberration is

$$\frac{j_1}{u_1}=-\frac{j_2}{u_2}=\frac{j}{u_1-u_2}.$$

Since in the alternative construction the sign of j_2 is altered, the condition that both types of lens are chromatically corrected is that $\frac{\delta\omega_2}{\omega_2}$, $\frac{\delta\omega_1}{\omega_1}$, $\frac{\delta\omega_2'}{\omega_2'}$ are in arithmetical progression. More generally, if this condition is satisfied, the amount of chromatic aberration will be the same in both types of lens having the same exterior glass.

To solve the condition $\gamma+\gamma_1=0$ when ω_1 and ω_2 are given, we first make the right side of equation (1) homogeneous in j and j_2, using equation (2) for j_1. The result is

$$\frac{\omega_1}{1-\omega_1}\left(\frac{j_2}{j}\right)^3+\frac{j_2}{j}+\frac{1}{\omega_1-\omega_2}=0.$$

Writing

$$x=2\left(\frac{1-\omega_1}{3\omega_1}\right)^{1/2},\quad \sinh 3\theta=\frac{3}{(\omega_2-\omega_1)x},$$

the solution is

$$j_2=jx\sinh\theta,$$

and by (2)

$$1/j=1-\omega_1-(\omega_2-\omega_1)x\sinh\theta.$$

Constructional details for lenses of these types, when thicknesses are neglected, can therefore be obtained directly from common tables.

It will be noticed that the problem considered here is essentially the same as the determination of a cemented doublet corrected for parallel light incident initially on a plane surface.

§ 3. EXAMPLES OF APPLICATION

It may be of interest to add two numerical examples, one where the glasses do not differ greatly in their indices, and the other where the difference is large. Let $\omega_1 = 0{\cdot}600$, $\omega_2 = 0{\cdot}610$. Then $x = \frac{2}{3}\sqrt{2}$, $\sinh 3\theta = 225\sqrt{2}$,

giving $$\sinh\theta = 3\sqrt{2}, \quad j = \frac{25}{9}, \quad j_2 = \frac{100}{9},$$

and the curvatures are

$$\frac{8}{9}, \quad \frac{61}{18}, \quad -\frac{61}{18}, \quad -\frac{8}{9}.$$

The corresponding alternative form has

$$\omega_1 = 0{\cdot}600, \quad \omega_2' = 0{\cdot}590,$$

with curvatures

$$\frac{8}{9}, \quad -\frac{59}{18}, \quad \frac{59}{18}, \quad -\frac{8}{9}.$$

For chromatic correction

$$\frac{\delta\omega_2}{\omega_2} : \frac{\delta\omega_1}{\omega_1} : \frac{\delta\omega_2'}{\omega_2'} = 3:4:5.$$

The 25 per cent differences in these chromatic factors contrast markedly with differences in the indices of under 2 per cent.

For the second example take $\omega_1 = 0{\cdot}500$, $\omega_2 = 0{\cdot}600$. This gives $x = \frac{2}{3}\sqrt{3}$, $\sinh 3\theta = 15\sqrt{3}$, with the solution

$$\sinh\theta = \sqrt{3}, \quad j = \frac{10}{3}, \quad j_2 = \frac{20}{8}.$$

The curvatures are $$\frac{7}{6}, \quad 2, \quad -2, \quad -\frac{7}{6}.$$

In the associated solution for $\omega_1 = 0{\cdot}500$, $\omega_2' = 0{\cdot}400$ (corresponding to a higher refractive index than has hitherto been obtained in glass) the curvatures are

$$\frac{7}{6}, \quad -\frac{4}{3}, \quad \frac{4}{3}, \quad -\frac{7}{6},$$

and for chromatic correction

$$\frac{\delta\omega_2}{\omega_2} : \frac{\delta\omega_1}{\omega_1} : \frac{\delta\omega_2'}{\omega_2'} = 1:2:3.$$

The substantial difference between the curvatures of the inner surfaces of the two lenses is to be expected in view of the very great difference 0·2 in the ω values. The 20 per cent differences in these values are accompanied by 50 per cent differences in the dispersive ratios. The examples suggest that the former difference increases relatively more rapidly than the latter.

§ 4. ACKNOWLEDGMENT

The work described above has been carried out as part of the research programme of the National Physical Laboratory, and this paper is published by permission of the Director of the Laboratory.

REFERENCE

SMITH, T., 1945. *Proc. Phys. Soc.*, **57**, 543.

REFLEXION OF CENTIMETRIC ELECTROMAGNETIC WAVES OVER GROUND, AND DIFFRACTION EFFECTS WITH WIRE-NETTING SCREENS

By J. S. HEY, S. J. PARSONS and F. JACKSON,
Ministry of Supply

MS. received 19 *December* 1946

ABSTRACT. Difficulties in the operation of centimetric wave-length radar equipment at low angles of elevation (less than 10°) have led to a detailed consideration of the influence of ground reflexion. A technique is described for the determination of the relation between the signal strength of the echo from an isotropic reflector and the angle of elevation of the reflector. Measurements obtained over natural ground sites are shown to be in accordance with simple theoretical considerations. The effect of wire-netting artificial screening has been examined, the experimental results being in general agreement with those derived by theoretical treatment of diffraction using Sommerfeld's formula.

§ 1. INTRODUCTION

The wartime requirements for Army radar equipments to operate at angles of elevation of a few degrees gave rise to a number of siting problems. In order to avoid the clutter on the cathode-ray tube display of unwanted echoes from ground objects such as buildings or hills, which on high sites may return appreciable signals at many kilometres range, it was found desirable to place the radar set so that a crest within a few kilometres afforded screening from more distant ground objects. Artificial screening by wire netting was also introduced during the flying bomb attacks on London and S.E. England for radar sites where no suitable natural crests were available.

In addition to the requirement for eliminating the clutter of echoes from ground objects, it was essential to know the coverage of the radar set at low elevations, this being profoundly affected by reflexion and diffraction effects of the ground and screen surrounding the set. The present research was carried out between autumn 1944 and spring 1945 at the experimental station of the Army Operational Research Group, Ministry of Supply. The investigation was designed to determine the echo signal strength pattern in the vertical plane for an Army radar equipment operating at a wave-length of 10·7 cm. on typical natural sites both with and without artificial screening. An attempt is here made to relate the results to theoretical considerations of reflexion and diffraction.

§ 2. EXPERIMENTAL TECHNIQUE

The equipment consisted of the British Army set known as G.L. III (see plate 1), with some minor receiver modifications to give a suitable presentation for the measurements required in the present investigation. The transmitter radiated pulses of 1 microsecond duration with a peak power of about 200 kw. Separate transmitter and receiver paraboloids with apertures of 1·22 m. diameter mounted adjacently with foci at a height of 3·6 m. could be traversed together in bearing or elevation as required. Reflected pulses were displayed on the receiver cathode-ray tube as deflections of a linear range trace. The receiver output circuits were arranged so that the deflection amplitude on the cathode-ray tube was directly proportional to the received signal strength.

The investigation was carried out for vertical polarization (electromagnetic waves with the electric vector in a vertical plane). The transmitting aerial was fixed, but the receiving aerial was displaced alternately to the right and left so as to give a horizontal deflection of the electrical axis of the beam of 1·2° to each side. The received signals for the two positions were presented so as to appear side by side on a second cathode-ray tube, the amplitudes being equal when the paraboloid axis was directed towards the bearing of the reflector. This method was the standard radar technique for determination of bearing by this equipment and was used throughout the investigation to maintain the paraboloid axis on the bearing of the target.

The reflector from which the echoes were obtained consisted of a 0·6 m. diameter papier-mâché sphere, metallized with sprayed zinc, suspended from a tethered balloon at ranges of 2500 to 3000 m. Weather conditions, with frequent gusty winds, often presented a serious handicap to experimental work. The most satisfactory method of supporting the sphere was found to be by means of an "M" type balloon (a spherical fabric balloon of about 4 m. diameter) flying on a nylon cord. The balloon itself gave a signal about $\frac{1}{8}$ of that from the sphere; provided the balloon and sphere were close together, the slight beating of the signal due to this cause could be tolerated. By measurements of the pulse amplitude on the cathode-ray tube, and with the aid of several fixed receiver gain settings in simple ratios, the signal strength could be determined in terms of the "free-space" reflexion signal to an accuracy of roughly 5%. The free-space signal was obtained from observations at an angle of elevation of about 15°, where ground interference effects were negligible.

The transmitter and receiver paraboloid axes were set at a fixed elevation of a few degrees and the height of the spherical reflector was varied at a rate of roughly 15 m. per minute between ground level and about 500 m. The elevation of the reflector was measured by means of a theodolite placed close to the equipment. Readings were taken at intervals of two seconds both of angle of elevation of the reflector and of echo signal amplitude on the cathode-ray tube. The range of the echo was also recorded so that, where necessary, the readings of echo signal amplitude could be normalized to a constant range by the inverse-square law.

§ 3. ELEMENTARY THEORETICAL CONSIDERATION OF GROUND REFLEXION

The manner in which the signal strength echoed from an isotropic reflector varies with the angle of elevation of the reflector for a given bearing, and elevation of the paraboloid axes will here be referred to as the directional sensitivity * of

---

* In the Services and elsewhere this has often been loosely referred to as the vertical polar diagram of the equipment.

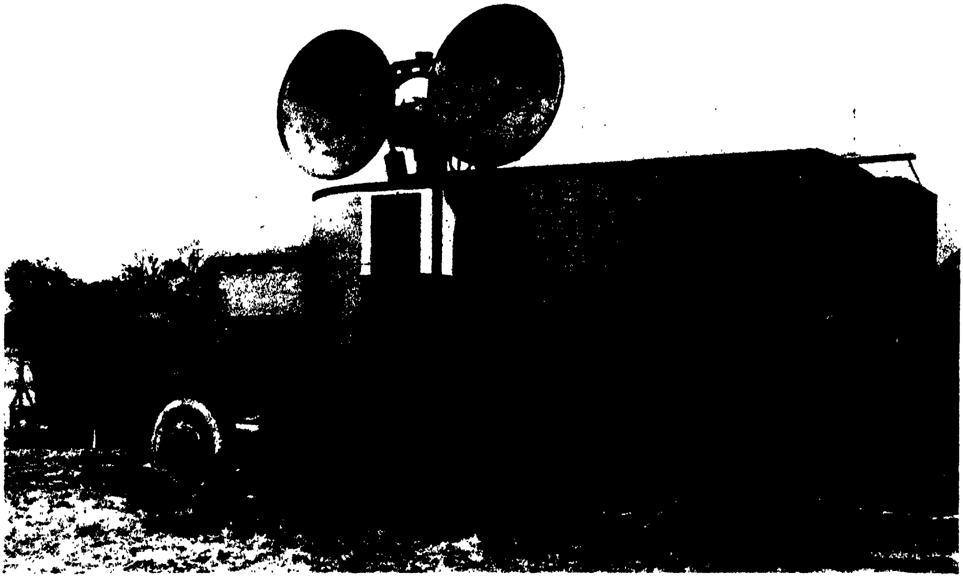


Plate 1.

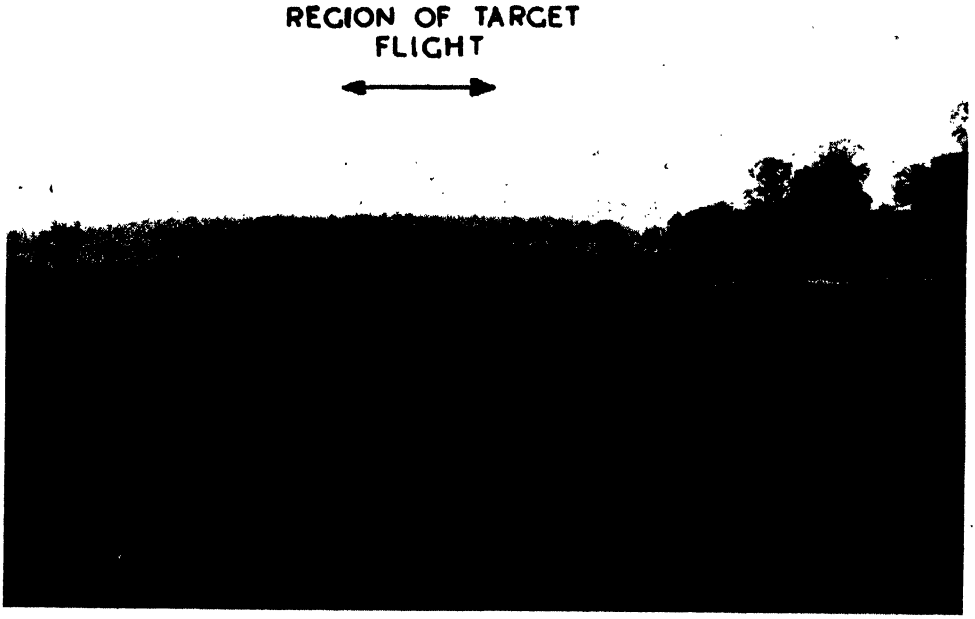


Plate 2.

the radar equipment for the vertical plane considered. Where ground reflexion is present, the directional sensitivity pattern will depend on the configuration and reflexion coefficient of the ground, the position of the radar aerial, and the free-space directivity factor of the aerial. In the case of the transmitting aerial, we define the free-space directivity factor as the relation between direction and radiated field strength, scaled to a maximum of unity; for the receiving aerial, it is the relation between the direction of an incident plane wave of given amplitude, and the received signal strength, again scaled to a maximum of unity. Suppose the free-space directivity factor of the transmitting aerial in the vertical plane is $F(\theta)$, where θ is the angle between the axis of the paraboloid and the direction of the target, then the field strength at a reflector, elevation α, illuminated by a paraboloid tilted at elevation ϕ, compared with a maximum free-space value of unity, is given by

$$F(\theta)_{\alpha-\phi}+Re^{i\psi}F(\theta)_{-\alpha-\phi},$$

where the expression represents the addition of the direct and reflected waves, R being the reflexion coefficient (which may be a complex quantity), ψ the phase difference due to the path difference between the direct and reflected waves, and $i=\sqrt{-1}$. If the reflector is isotropic, the directional sensitivity of the radar equipment is the product of the above expression and one similarly derived for the receiving aerial.

Figure 1 shows the calculated pattern of directional sensitivity in a vertical plane for the G.L. III aerials, at a height of 3·6 m., directed at grazing incidence to ground having a horizontal plane surface with assumed values of the reflexion coefficient of $R=-1\cdot0$ and $R=-0\cdot5$ respectively. The diminution in amplitude of the lobes as the elevation increases is due to free-space directivity factor, $F(\theta)$, of the aerials. The multiple lobe structure is due to ground reflexion interference,

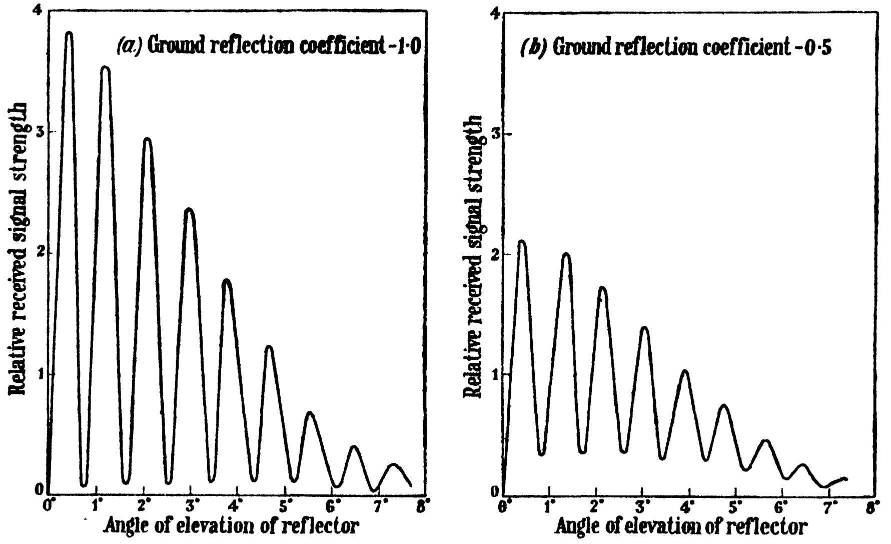


Figure 1. Calculated directional sensitivity patterns for G.L. III.
Aerial height 3·6 m. above horizontal plane ground. Radar axis at 0·8° elevation.

The elevations of the maxima are given approximately by

$$\sin\alpha = (n - \tfrac{1}{2})\lambda/2h,$$

and of the minima by

$$\sin\alpha = (n-1)\lambda/2h,$$

where α is the angle of elevation, λ the wave-length, h the aerial height above the plane, and n is an integer. For $\lambda = 10{\cdot}7$ cm. and $R = 3{\cdot}6$ m. the corresponding values of α are as follows:—

Maxima (°) 0·4, 1·2, 2·0, 2·8 ....
Minima (°) 0, 0·8, 1·6, 2·4 ....

If the plane has a slope of angle of elevation β in the direction of the reflector, then the above angles will be increased by β in each case. For the case of a distant slope, differing from that of the near ground, simple image theory leads to substitution of an effective height for h in the above equations for those reflexions that take place on the distant slope (provided at least about half the first Fresnel zone lies on the slope). This effective height is the perpendicular height of the aerials above the plane containing the distant slope. In such cases, however, where the ground surface is discontinuous, multiple reflexions often add further complexity.

Although natural ground is generally far from flat, certain deviations can be tolerated before we may expect serious changes in the shape of the directional sensitivity patterns shown in figure 1. Provided the region around the point of reflexion (as defined by geometrical optics) is flat for an area of the order of half the first Fresnel zone, the reflected wave will be of the same order as that for an infinite reflecting plane. Ground may be said to be flat if the surface irregularities do not cause path differences exceeding a small fraction of a wave-length. This may be expressed in the form

$$H\sin\alpha \ll \lambda/2,$$

where H is the height of the surface irregularity above the mean level. For example, the measurements of Ford and Oliver (1946) indicate that, for $H\sin\alpha$ approximately equal to $0{\cdot}2\lambda$, the value of the reflexion coefficient is about 0·5.

The dimensions of the first Fresnel zone may be calculated as follows. In figure 2, let A represent an aerial at height h above a plane reflecting surface. B is the point of reflexion, such that the angle of incidence is equal to the angle of reflexion, and CD is the reflected wave front. Then the outer limit D of the first zone is determined by

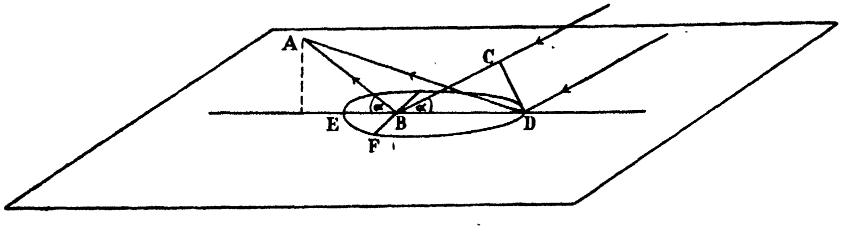


Figure 2. Diagram to illustrate calculation of first Fresnel zone.

$$AD - (AB + BC) = \lambda/2.$$

This can be approximately expressed by

$$BD=\frac{\lambda\cos\alpha+\sqrt{\lambda^2\cos^2\alpha+4h\lambda\sin\alpha}}{2\sin^2\alpha}.$$

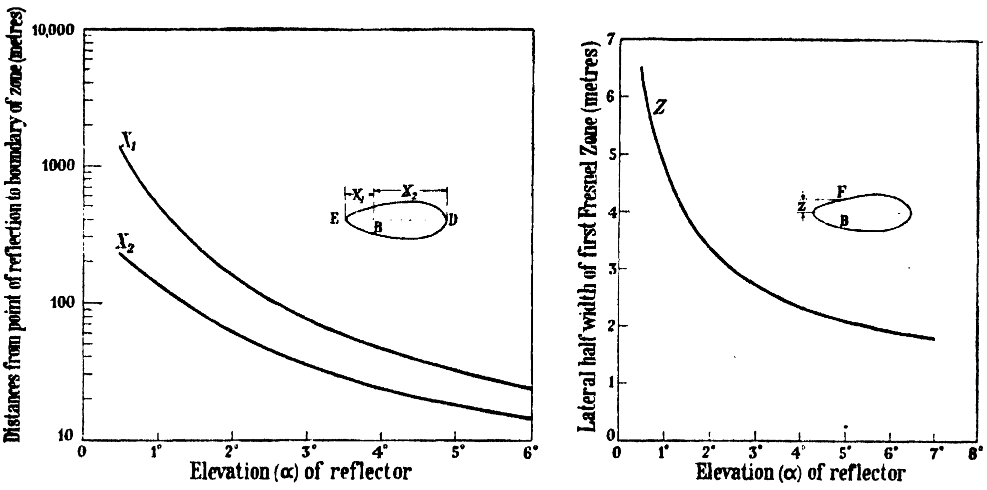
Similarly, the inner limit E is approximately given by

$$BE=\frac{-\lambda\cos\alpha+\sqrt{\lambda^2\cos^2\alpha+4h\lambda\sin\alpha}}{2\sin^2\alpha},$$

and the lateral half-width is approximately given by

$$BF=(h\lambda/\sin\alpha)^{1/2}.$$

The relations between zone size and angle of elevation are given graphically for $h=3{\cdot}6$ m. and $\lambda=10{\cdot}7$ cm. in figure 3, where BD, BE, BF are represented by X_1, X_2, and Z. Two features are of particular note in practical considerations of ground sites. Firstly, the zones are narrow compared with their length, so that the nature of the interference pattern on a given bearing will be determined by the characteristics and slope of the ground in that direction, and will be independent of the ground more than a few degrees away from this bearing. Secondly, as the elevation of the reflector increases, the length of the first zone diminishes and the point of reflexion approaches the equipment. It may be shown that the zones are truly elliptical in shape.



$$X=\pm\frac{\lambda\cos X+\sqrt{\lambda^2\cos^2\alpha+4h\lambda\sin\alpha}}{2\sin\alpha}.$$

$\lambda=10{\cdot}7$ cm., $h=3{\cdot}6$ m.

$$Z=\sqrt{\frac{\lambda h}{\sin\alpha}}.$$

$\lambda=10{\cdot}7$ cm., $h=3{\cdot}6$ m.

Figure 3. Relations between dimensions of first Fresnel zone and angle of elevation of the reflector.

In order to attempt to apply the above considerations to natural sites, a survey of the ground was made in the directions in which echo signal strength measurements were taken. From the vertical profile, a mean point of reflexion could be estimated for any given value of the reflector elevation α. If the ground around this point was flat (within the specification discussed above) over an extent not less than half the first Fresnel zone, a reflexion could be expected in accordance with simple theory of reflexion from a plane surface.

In the following section a few typical results, obtained at four sites in Richmond Park, are described For two sites (I and II), the above conditions for simple reflexions are largely satisfied; two other sites (III and IV) represent more complex cases. In all cases the results for the echo signal amplitude are expressed in terms of the maximum free-space value.

§4. EXPERIMENTAL RESULTS FOR REFLEXION AT NATURAL SITES

Site I. A photograph of the site is shown in plate 2, and figure 4 gives the measured contour of a vertical section of the ground in the direction in which radar observations were taken. (In the ground contour diagram it should be noted that height variations and angles of elevation appear exaggerated, due to the difference between the height and range scales.)

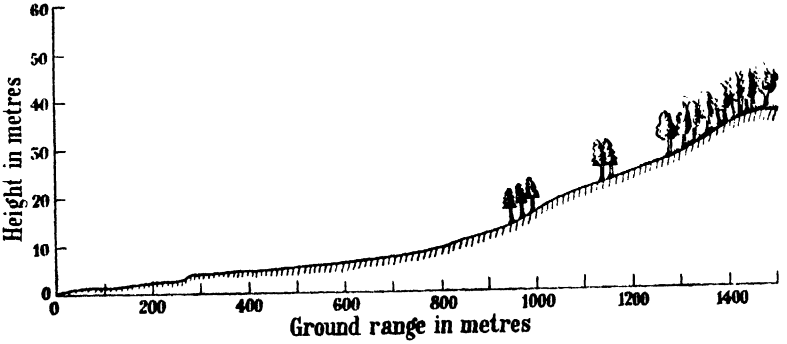


Figure 4. Vertical Section of Site I.
(Note that height and range scales are different.)

Simple theoretical considerations according to the principles outlined above indicate that the ground up to approximately 800 m. from the radar equipment could be considered for the most part flat, with a slope of 0·5°. The angle of elevation of the crest was 2·2°, and for elevations above this value the points of reflexion for a plane surface are within 300 m. Thus we may expect interference maxima at angles of 0·5° greater than those for a horizontal plane reflecting surface. The theoretical lobe maxima will therefore be at 2·5°, 3·3°, 4·1°, etc. The higher lobes are less easily predicted for the following reasons. As the elevation increases, the size of the first Fresnel zone becomes less, and hence departures from the mean slope of 0·5° may occur over the area of the zone; irregularities within the zone also become more important. Further, at the higher angles of elevation, the lobes are more likely to be influenced by secondary reflexions from the crest slope (ranges beyond 800 m. from the equipment), although such reflexions can only have a minor effect owing to the considerable variations in slope and the wooded nature of the ground in this region.

Measurements of the signal reflected from the sphere (at about 3000 m. range) were made with the axes of the radar paraboloids at fixed elevations of 0°, 3° and 5° respectively. Smooth curves drawn through the graphed results are shown in figure 5. In the case of the paraboloid axes at 3° elevation, the points representing the individual measurements have also been marked; this has been chosen as a typical example to demonstrate the scatter in the observations. It will be seen that there is general agreement between the expected positions of the interference

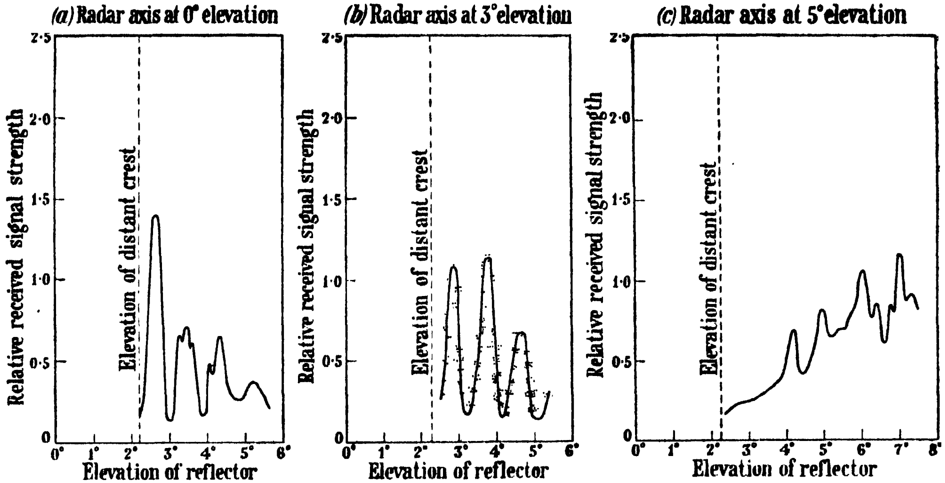


Figure 5. Directional sensitivity pattern at site I.

lobes and those measured experimentally. The increased magnitude of the first observable lobe, when paraboloids are set with their axes at 0° elevation, is notable and indicates a high value for the reflexion coefficient.

Site II. The vertical ground section for the site is shown in figure 6. The ground from about 100 m. to 300 m. from the radar equipment may be regarded as having an almost uniform mean slope of about 0·9°, the effective height of the aerial with respect to this region being approximately 4 m. The measured vertical directional sensitivity pattern is shown in figure 7. The small lobe at 2·3° could be accounted for by the lobe with predicted maximum at 2·0°, which would be considerably screened by a distant heavily-wooded crest at 2·1° elevation (not shown in figure 6). The observed maxima at 2·8° and 3·6° fit well with the values of 2·75° and 3·5°, which would be expected from the above considerations of the site contour. The discontinuities of the ground within about 100 m. make the lobe structure for higher angles of elevation more difficult to estimate.

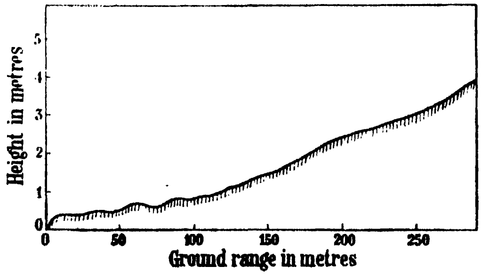


Figure 6. Vertical section of site II. (Note that height and range scales are different.)

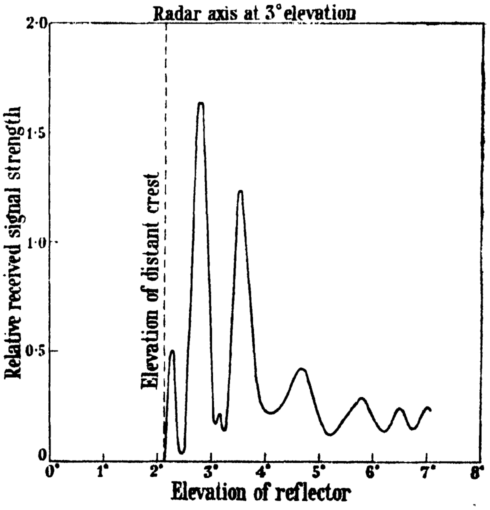


Figure 7. Directional sensitivity at site II.

Site III. The vertical section for this site is illustrated by figure 8, and the measured echo strength pattern in figure 9 shows a strong lobe system extending to high angles. This might be expected from the steepness of the gradients; further, the ground slope changes from concave to convex before attaining a crest only some 500 m. away, and no simple prediction of lobe positions or magnitudes is possible owing to the complexity of the contour.

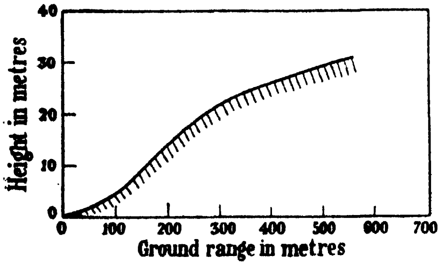


Figure 8. Vertical section of site III. (Note that height and range scales are different.)

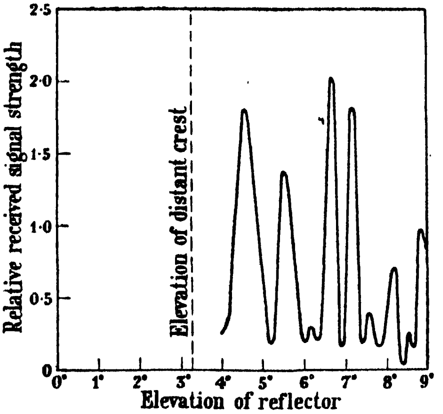


Figure 9. Directional sensitivity pattern at site III. Radar axis at 4° elevation.

Site IV. Consideration of the vertical section at this site, given in figure 10, indicates that double or triple reflexions may occur. A complicated relation between echo strength and reflector elevation might be expected, therefore, and this is shown to be the case in figure 11. As for the previous site, no simple prediction of lobe positions or magnitudes is possible.

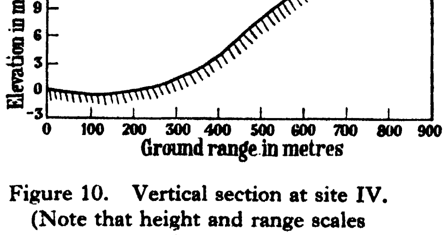


Figure 10. Vertical section at site IV. (Note that height and range scales are different.)

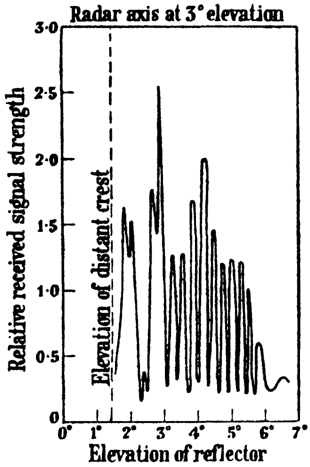


Figure 11. Directional sensitivity pattern at site IV.

§ 5. DIFFRACTION EFFECTS WITH ARTIFICIAL SCREENS

On certain radar sites where there was no natural crest to afford screening from unwanted echoes from ground objects, wire-netting screens of 22 S.W.G. galvanized iron wire and 1·4 cm. mesh were erected after a preliminary trial had indicated their effectiveness for this purpose. These screens were placed at a distance of about 50 m. from the radar equipment, the top edges of the screens being at the lowest elevation compatible with adequate reduction of the ground clutter.

Measurements of the effect of such screens on the directional sensitivity pattern, described below, demonstrated a general agreement with the following theoretical treatment of diffraction using Sommerfeld's solution. The Sommerfeld formula for the diffraction of electromagnetic waves by a semi-infinite perfectly reflecting plane screen was applied to the two parallel edges of a screen of finite width.

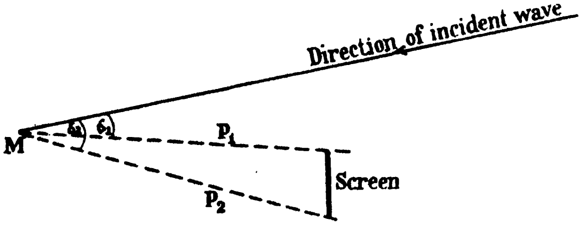


Figure 12. Diagram to illustrate calculation of diffraction effects.

We are here concerned with diffraction through small angles, and in this case, with the notation in figure 12, the received amplitude and phase angle of the wave at a point M, as compared with the value if no screen were present, are given by the complex amplitude

$$A=\frac{1}{\sqrt{2}}e^{\frac{-i\pi}{4}}\left\{\int_{-\infty}^{Z_1} e^{\frac{-i\pi}{2}v^2}\,dv+\int_{Z_2}^{+\infty} e^{\frac{-i\pi}{2}v^2}\,dv\right\},$$

where

$$Z_1=2\sin\frac{\delta_1}{2}\sqrt{\frac{2p_1}{\lambda}},\qquad Z_2=2\sin\frac{\delta_2}{2}\sqrt{\frac{2p_2}{\lambda}}.$$

The values of these integrals are readily obtainable from Cornu's Spiral.*

Since a paraboloid aerial system is used, and not a point aerial, the following artifice was adopted. Consider first a plane wave received by the paraboloid in free space. The received signal was derived from the wave at the circular aperture of the paraboloid by dividing the aperture into horizontal strips. Let S be the area of a strip expressed as a fraction of the whole aperture. The amplitude and phase of the wave received by the strip, compared with those obtained if the aperture was rotated about its centre so as to be parallel with the wave-front, may be expressed by $Se^{i\phi}$, where ϕ is the phase angle at the centre of the strip with respect to the phase at the centre of the paraboloid aperture. The resultant signal received is given by the complex summation for all the strips.

If now a screen is introduced in front of the aperture, the diffracted signal received by a strip becomes $ASe^{i\phi}$, where A is given by the diffraction formula above. A similar treatment may be carried out for the waves reflected from the ground. In the present case, where the screen is near the radar set, the only important ground reflexion is that which takes place from the ground on the far side of the screen. The resultant received signal, compared with the maximum free-space value, can then be determined by combining the contributions of all the strips for both direct and reflected waves with due regard to their relative phases. By the reciprocity theorem the same result applies for transmission as for reception. This method has been used, assuming flat ground with a reflexion

---

* For numerical values, see, for example, Jahnke and Emde's *Tables of Functions*.

coefficient of $-1{\cdot}0$ and arbitrarily choosing the number of strips as 5, to derive the theoretical echo strength for various elevations of a spherical reflector in terms of maximum free-space signal.

The experimental measurements with screens were made on Site I (figure 4), which was the flattest available site in the Richmond Park trials ground. It should be noted that a site of this type, with a natural crest at 2·2° elevation, does not require wire netting to provide screening from the ground clutter. It was possible to demonstrate, however, a satisfactory agreement between experiment and theory by measurements at elevations above that of the natural crest. This is shown, for example, by figure 13, in which measurements and theoretical calculations are compared in the case of a screen of 4 m. vertical depth.

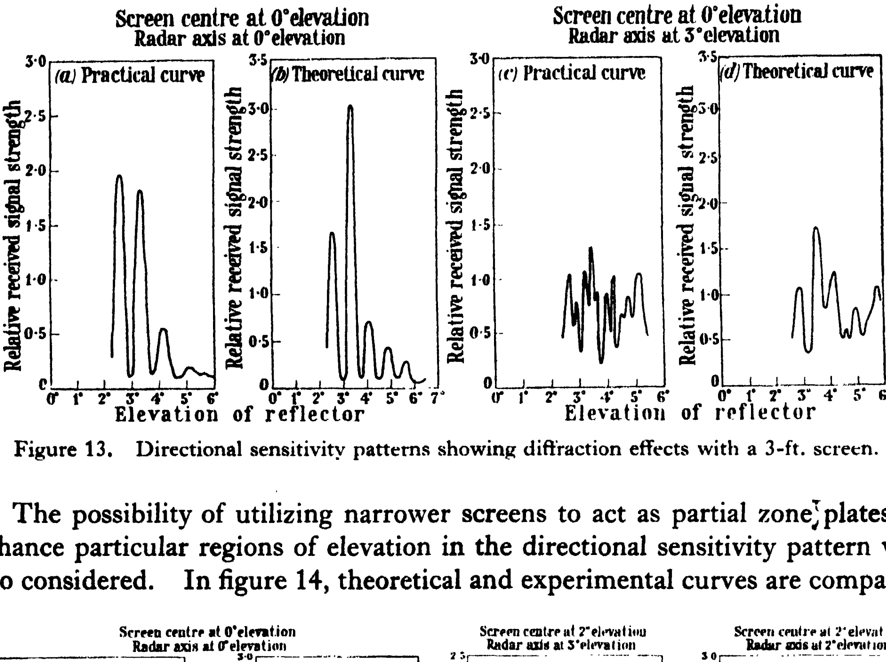


Figure 13. Directional sensitivity patterns showing diffraction effects with a 3-ft. screen.

The possibility of utilizing narrower screens to act as partial zone plates to enhance particular regions of elevation in the directional sensitivity pattern was also considered. In figure 14, theoretical and experimental curves are compared

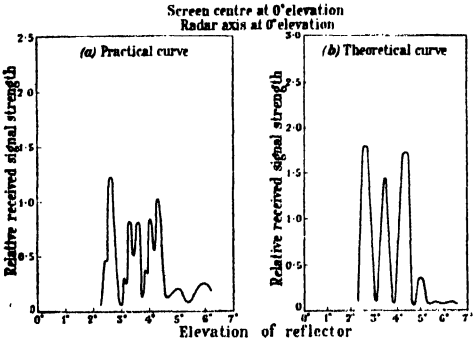


Figure 14. Directional sensitivity pattern showing diffraction effects with a 7½-ft. screen.

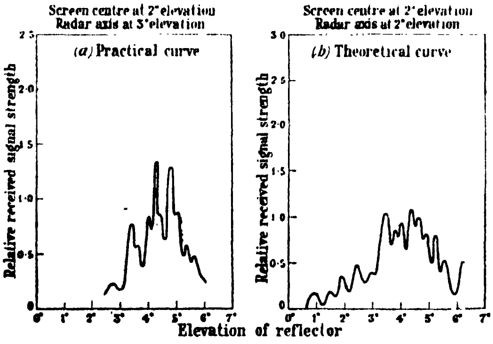


Figure 15. Directional sensitivity pattern showing diffraction effects with a 12-ft. screen.

for screens of vertical depth 1 m. in a position chosen on zone-plate principles so as to enhance the lobes from 2·4° to 4·4° elevation. Between these angles, the screen cuts off approximately the second Fresnel zone in the lower half of the direct wave front and in the upper half of the reflected wave-front. In figure 15, the experimental and theoretical results are shown for a screen of 2·25 m. designed similarly to enhance the lobes between 3·7° and 5·8° elevation by cutting off the second, third and fourth zones. Again a general agreement in form between the measurements and theoretical expectations is apparent.

§ 6. CONCLUSIONS

The investigation described above has demonstrated a method for the determination of the directional sensitivity pattern in a vertical plane for radar equipment of centrimetic wave-length and has shown that the reflexion and diffraction effects can be predicted on many natural sites to a first approximation by simple theoretical considerations. The varying configuration and nature of the ground of natural sites is often such that precise measurements of magnitude and phase of the reflexion coefficients are of little practical value since a solution involving detailed variations would be too complex to solve. It is of practical interest therefore to see how well simple considerations can often give a satisfactory forecast of the directional sensitivity pattern.

The results in figures 5 to 10 suggest, by the depth of the minima and amplitude of the maxima, that the reflexion coefficients for rough pasture land for vertical polarization has a high value for the elevations considered, namely 0° to about 8°. It may be noted that the experiments of Ford and Oliver (1946) showed that, for vertical polarization, grass can give higher values of reflexion coefficient than does bare ground.

The wire-netting screen used in the diffraction experiments was far from opaque, as later experiments revealed. Nevertheless the experimental results show satisfactory general agreement with the simple application of diffraction theory for perfectly reflecting screens. It is also seen that practical realization of zone-plate effects may be obtained by suitable choice of screen dimensions.

§ 7. ACKNOWLEDGMENTS

The above investigation was carried out as part of the programme of the Operational Research Group, Ministry of Supply, J. T. G. Milne and G. S. Stewart taking part in the early experimental investigations with wire-netting screens. The authors are indebted to the Chief Scientist, Ministry of Supply, for permission to publish this communication.

REFERENCE

Ford and Oliver, 1946. *Proc. Phys. Soc.*, **58**, 265.

RADAR OBSERVATIONS OF METEORS

By J. S. HEY and G. S. STEWART,
Ministry of Supply

Read 31 *January* 1947 ; *MS. received* 6 *March* 1947

ABSTRACT. (i) An investigation of short-duration radio echoes observed at 4 to 5 metres wavelength in the neighbourhood of the E region of the ionosphere is described. Observations by vertical beam radio equipments showed that the echoes occurred most frequently at a height of about 95 km. Marked directional characteristics were revealed by the use of equipments with oblique beams, the diurnal variations being different on different bearings.

(ii) Analysis of the results has indicated the close link in characteristics of these echoes and meteors, and has thus confirmed the suggestion of meteoric origin which has been made by some previous workers. A detailed correlation is demonstrated in the present investigation and a method is described for the determination of the radiants of the most active meteor streams. Further, a determination of geocentric velocities was made possible by the introduction of improved photographic techniques for the observation of the Giacobinid meteor shower in October 1946.

§ 1. INTRODUCTION

SINCE October 1944 we have used Army radar equipment operating on a wavelength of about 5 metres to investigate the transient ionospheric echoes obtained at heights around 100 km. The general occurrence of these short-duration echoes at frequencies exceeding the critical frequencies for either the normal or abnormal E layers was noted by Appleton, Naismith and Ingram (1937) during their Polar Year observations of 1932–33. Schafer and Goodall (1932), who worked in collaboration with Skellett (1932) in an investigation of meteors as a source of abnormal E-region ionization, also recorded them as a specific feature of the 1931 Leonid shower. Skellett (1935) was able to show that in certain cases sudden increases in abnormal E-layer ionization coincided with the passage of visible meteors. Further, Skellett (1938) considered that Eckersley's observations (1937) of the characteristics of the transient echoes indicated the ionization produced by the passage of meteors through the upper atmosphere as the origin of these echoes. Subsequent investigations of the echoes were made by Appleton and Piddington (1938), who determined the range distribution and reflexion coefficients, and Eckersley (1940), who analysed the durations and made some observations of the diurnal variation in rate of occurrence of the echoes. Eckersley and Farmer (1945) have recently made detailed measurements of the polarization and direction of the reflexions received; they thought that their results did not conform with the meteoric hypothesis. Appleton (1945), however, in his Kelvin Lecture, considered that the evidence so far available had strongly suggested the meteoric origin.

The developments in radar techniques which have taken place during the war have made it possible for more exact determinations of some of the characteristics of these transient echoes. We have recently published (1946) a brief description of some of the results of our investigations at wavelengths between

4 m. and 7 m. Ferrell (1946) indicates that measurements are also being made in America on wavelengths of this order. Such wavelengths are considerably shorter than those used in the early work to which reference has been made above. Our investigations, which are here described in detail, provided conclusive proof of the meteoric origin of the echoes by a comparative study of their characteristics and the properties of meteors. The influence of the orientation of the meteor trail * in determining whether or not a reflexion may be obtained from it was first pointed out by Pierce (1938, 1941) and has been found to be a factor of particular importance in the interpretation of our results.

§2. INITIAL OBSERVATIONS, OCTOBER–NOVEMBER 1944

Transient ionospheric echoes were observed at wavelengths of 4 to 5 metres on certain long-range Army radar equipments with elevated beams during the latter of part 1944.† A chain of 12 of these radar sets was deployed by A.A. Command, and the authors were directly concerned in an advisory capacity, both in the planning of the system and the investigation of its operational performance. The transmitters each radiated approximately 500 pulses per second with a pulse duration of about 3 microsecs. and a peak power of 150 kw. The aerial systems provided elevated beams with axes around 45° to 55° elevation, some of the equipments having stacks of four dipoles and others single Yagi aerials. The receiver time base was extended to 140 km. and the characteristics of all echoes exceeding 2 seconds in duration were noted by direct visual observation of the cathode-ray tube display.

An analysis of the transient ionospheric echoes recorded between 16 October and 19 November 1944 was made by E. B. Britton, who was working in collaboration with the authors. The analysis showed that of a total of 348 echoes the average duration was about 13 secs. The mean ratio of signal to noise was approximately 4, the peak values in amplitude appearing at the onset of the echo. The average initial range was about 124 km. Assuming the mean elevation of observation to be that of the maximum of the radio beam, the mean heights were calculated to be approximately 93 km. on first appearance and 91 km. on disappearance. The analysis also revealed the interesting feature that although there was a considerable overlap in the coverage of the stations it was rare for an echo to be observed simultaneously by more than one station. Further, a diurnal variation in frequency of occurrence was recorded, the maximum being around sunrise and the minimum after sunset.

§3. INVESTIGATIONS FROM JUNE 1945–JUNE 1946

The above series of observations was incidental to an operational watch maintained for other purposes. At the end of the war it became possible through the cooperation of A.A. Command to utilize their Army radar facilities for

---

* The terms " train ", " streak ", and " trail " have all been used by various writers, sometimes with slight distinctions in meaning (see, for example, Herschel (1911)). In this report we shall refer to the column of glowing gas formed by the passage of a meteor through the upper atmosphere as a " trail ".

† P. E. Pollard informs us that such echoes were observed at these wavelengths in 1937–8 during the development of Army Gun-laying Radar Equipment by the Ministry of Supply. It was not until 1944, however, when this type of equipment was modified to work at very long range, that the echoes were of practical significence in Army Radar operation.

experiments designed to elucidate more specifically the characteristics of these short echoes. To this end a watch was set up involving five radar stations, two with vertical radio beams at Aldeburgh and three with inclined radio beams at Richmond, Aldeburgh, and Walmer respectively. These stations were manned and organized by A.A. Command, while the authors were responsible for the scientific direction of the experiments and the analysis of results.

At the end of July 1945, A.A. Command were unable to continue to provide the personnel to man the five stations. Their participation ceased, therefore, except for some further valuable assistance in the reading of the photographically recorded films, and for a short trial in October with the vertical beam sets at Sheerness. The research was therefore continued on a more limited scale in Richmond Park by the authors, with assistance from other Operational Research Group personnel.

In the following account we begin by describing in chronological order the investigations with the vertical beam stations. During the first few months many important data were collected although no definite conclusion as to the cause of the echoes appeared warranted. The subsequent development of the investigation, however, revealed a significant correlation with meteors.

A. *Observations with two vertical beam equipments, June, July and October* 1945

During June, July, and part of October 1945, a comparison of performance was made for two equipments with beams directed vertically upwards. The polar diagrams showing the variation of radar signal power sensitivity with direction are given in figure 1. The purpose in having two stations was to compare their performance both with similar and varied conditions of operation. In the first few weeks they were operated independently at frequencies of 73 Mc./s. ($\lambda = 4{\cdot}1$ m.) and 55 Mc./s. ($\lambda = 5{\cdot}4$ m.); but subsequently both were maintained at 73 Mc./s. and the effects of varying polarization and of using a common transmitter instead of two independent ones were investigated.

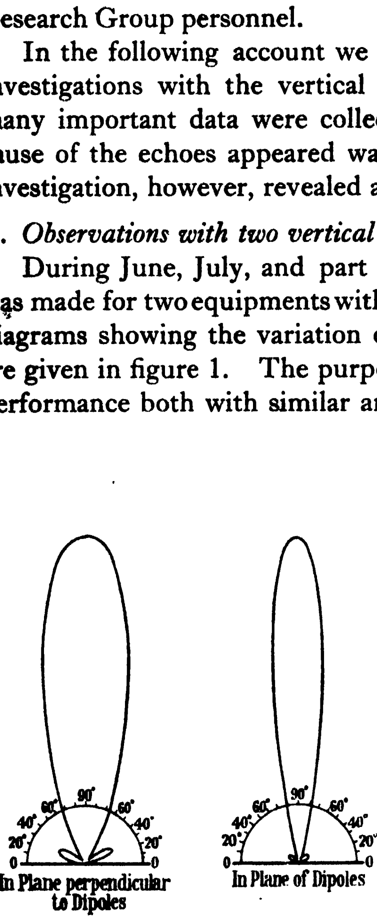


Figure 1. Polar diagram of radar signal power sensitivity. (Vertical-beam equipment A 1, A 2.)

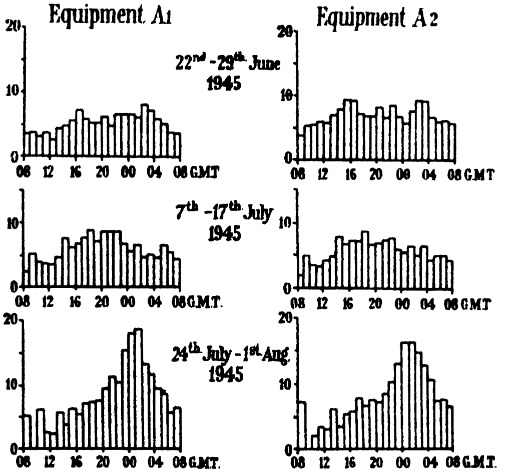


Figure 2. Diurnal variations in hourly rate of occurrence of echoes for vertical-beam equipments. Ordinate = mean hourly rate. Abscissa = time G.M.T.

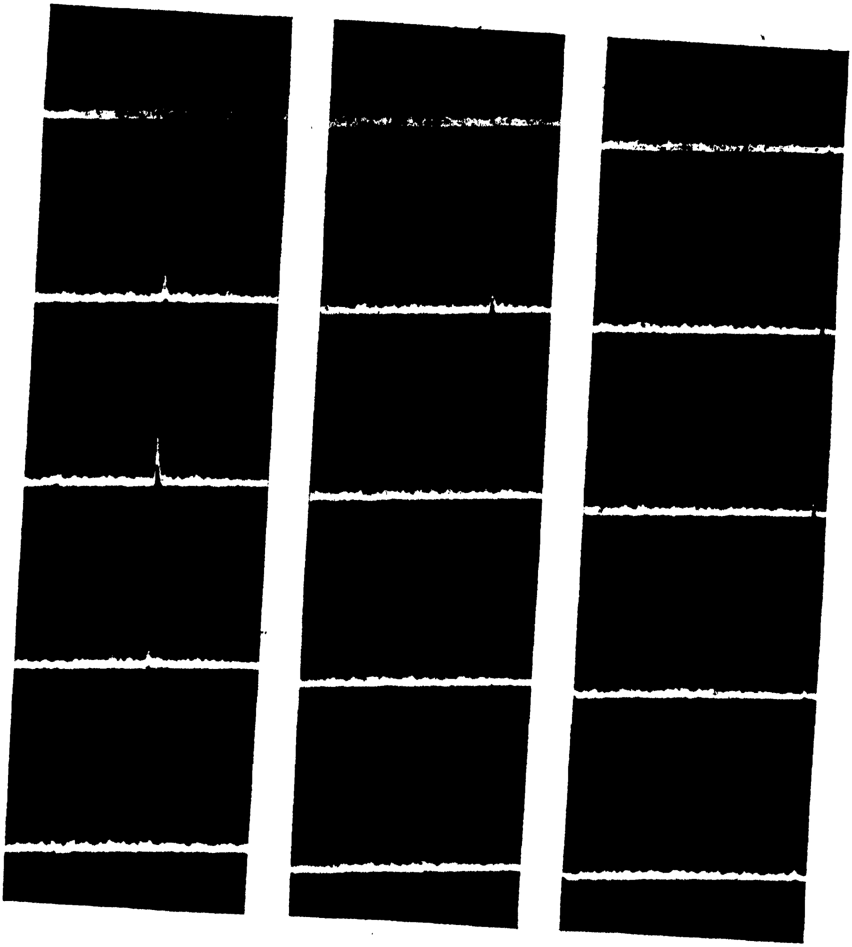


Plate 1.

Ciné photographs of transient ionospheric echoes presented as deflections of a linear range trace (abscissae, 80 to 125 km.); ciné speed—16 frames per second, sequence downwards.

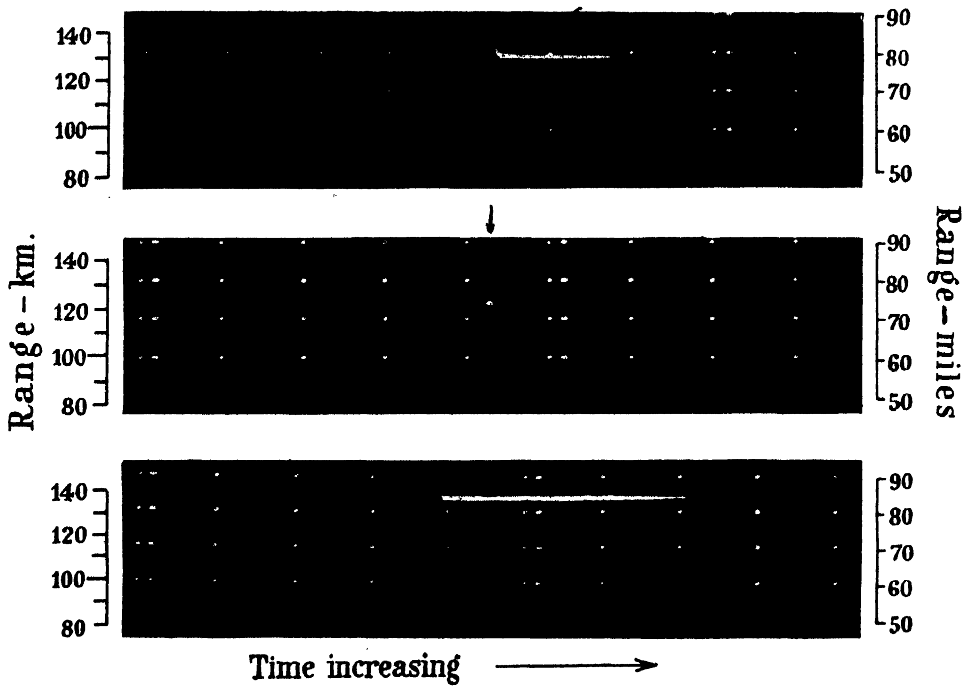


Plate 2.

Photographic records of transient ionospheric records presented as brightness modulation of range trace (ordinates, 80 to 140 km.) on continuously moving film. Range calibration marks appear at six-second intervals.

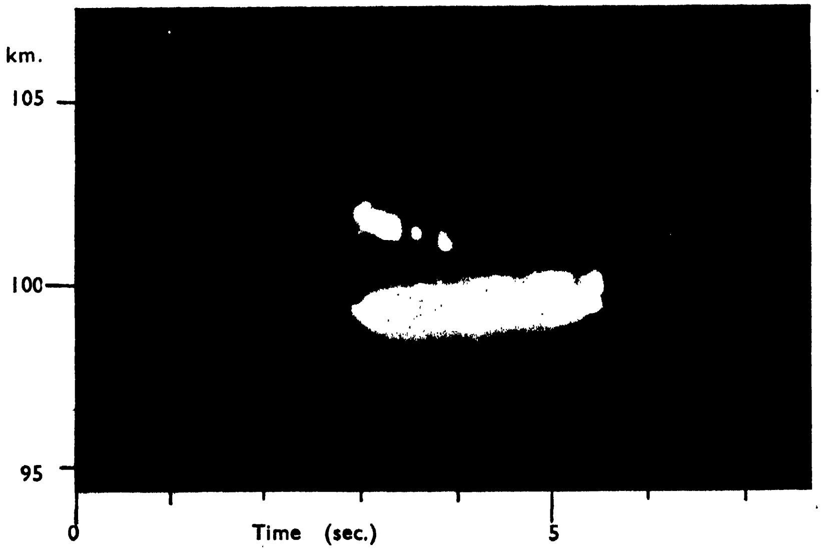


Plate 3.

Photographic record of a transient ionospheric echo during the Giacobinid meteor shower 1946.

A comparison was also made between visual recording of echoes displayed on the cathode-ray tube as deflections of a linear time base, and photographic recording of a brightness-modulated time base on a 16 mm. film moving at a speed of 0·5 mm. per sec. In the latter case, echoes were recorded on the film as spots if the echo duration was short, or as lines if the duration was long enough for the film to have moved appreciably. Ciné photographs of typical echoes presented as deflections on a linear time base, taken at 16 frames per second, are shown in plate 1. These illustrate the appearance of the time base to the operator making visual observations of the cathode-ray tube display. An example of an echo of very short duration is given in plate 1 (*b*) in which the echo can be seen during only one frame, and its duration must hence be about $\frac{1}{16}$ second, or less. Typical echoes obtained by photographic recording of the brightness modulated trace are shown in plate 2, which includes examples of echoes lasting several seconds.

We shall first consider the results for the three periods in table 1, for each of which the total number of echoes observed visually exceeded 1000. We refer to the two equipments as A 1 and A 2 respectively.

Table 1

Period	Details
22–29 June 1945	A 1 on 73 Mc./s., A 2 on 55 Mc./s. Visual recording.
7–17 July 1945	A 1 and A 2 on 73 Mc./s,, with common transmitter. Visual and photographic recording.
24 July–1 Aug. 1945	A 1 and A 2 on 73 Mc./s., with separate transmitters. Visual and photographic recording.

Graphs showing how the number of echoes per hour recorded visually varied with the time of day are given in figure 2 and the range distribution is shown in figure 3. These graphs demonstrate substantial agreement between the two stations, although the hourly rate for the 55 Mc./s. station, A 2, in the first period, is higher than that of the 73 Mc./s. station, A 1. Certain differences between the three periods are apparent in the mean diurnal variations of rate of occurrence, most particularly the pronounced peak in hourly rate which occurs around midnight in the third period; the interpretation of this peak is discussed later in this report (§ 4 *c*). The range distributions show no marked changes for the three periods. Since the radio beam is fairly narrow with its axis vertical, these represent approximate height distributions; thus the height of the region in which the maximum number of echoes occur is around 97 km. for each period.

Despite the fact that the two sets were on the same site with almost identical radar coverage, and that the form of the results reveals general similarity, the number of simultaneous observations recorded visually at the two stations amounted to slightly under 50% of the echoes seen by either station. This is explicable by the fact that the greater number of echoes were of momentary duration and small signal strength. As an example, the ratio of numbers of echoes of duration less than 1 second to those of greater duration was 3·4 : 1, while the ratio of the number of signals exceeding twice noise power to those exceeding four times noise power was 2·2 : 1. It is understandable therefore that observers might be especially liable to miss a proportion of echoes owing to the preponderance of echoes of short duration and small signal amplitude.

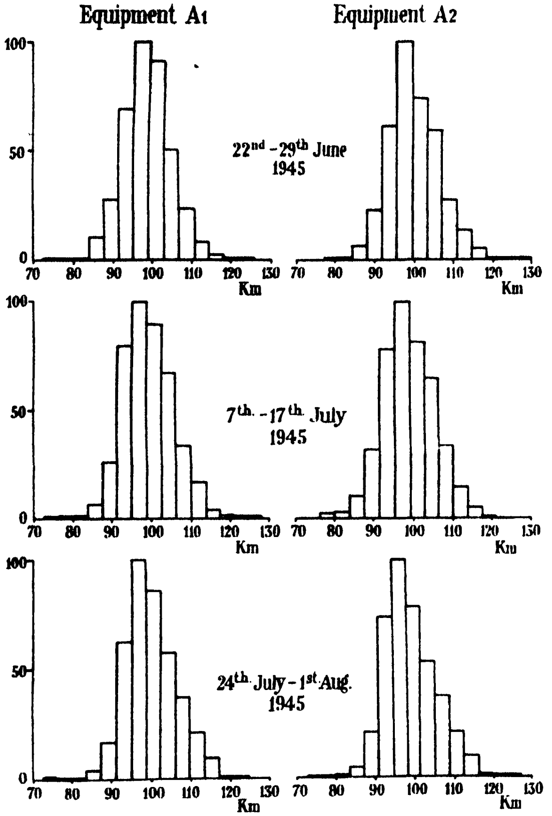


Figure 3. Range distribution of echoes for vertical-beam equipments.
Ordinate = relative number.
Abscissa = range.

Operation of the photographic recording of the brightness-modulated time-base was intermittent owing to technical faults, but the results indicated an increase of about 20% in the total number of echoes and a slight increase in the proportion of coincidences. The efficiency of this system depended on careful adjustment and stability in the setting of trace brightness and of receiver gain, these factors determining the noise level, which appears as a speckled background on the film. Imperfections in the recording apparatus prevented the coincidence rate from being greater, but it must be noted that the difficulties of recognizing small, short-duration echoes on a noise background are such that 100% coincidences can never be expected. With the knowledge of the earlier experiences, special precautions were taken during a few days of photographic recordings obtained at Sheerness in early October 1945, and the coincidence rate then rose to 85%.

In addition to the three periods of observation tabulated above, several experiments on polarization effects were undertaken for periods of one or two days. Although these tests covered such limited periods, the results indicated that there was no significant change when the plane of polarization was turned through 90°, from East–West to North–South. Further, when the polarization of a transmitter was at 90° to that of the receiver the number of echoes fell considerably, showing that the polarization of the echoed field tended to be the same as that transmitted.

B. *Observations with vertical beam equipment, December* 1945–*June* 1946

It was not practicable, owing to staffing problems, to continue the watch further until early December 1945. It was resumed at that date with a single vertical beam station in Richmond Park, the plan being to investigate over a long period any correlation between frequency of occurrence of echoes and meteor showers. Owing to the limitations in available staff the watch was kept only between 0915–1200 hrs. and 1400–1630 hrs., G.M.T. daily; these daily observations were made over a period December 1945 to June 1946 inclusive. All recording was made by visual observation of the echoes on the cathode-ray tube. During the period April 20–22 inclusive, the time of the Lyrid shower, the watch was carried out at night together with a visual watch for meteors in the sky. Daily measurements of transmitter field strength and receiver sensitivity were made so as to exclude equipment performance as a possible variable affecting the observed rate of occurrence of echoes.

(i) *Record of mean hourly rate, December* 1945–*June* 1946. The correlations in hourly rate of occurrence of echoes with the main meteor showers during this period has been discussed in a previous publication by the authors (1946). It will suffice here to illustrate the results by figure 4, and to draw attention to the marked peaks in hourly echo rate corresponding to the Quadrantid meteor shower on 2—3 January, and the Lyrids, on 20–22 April. The correlation is shown in an even more striking manner by figure 5, in which the hourly rate of broad and multiple echoes is plotted. A prominent peak for a single day occurs at the time of the Quadrantid shower, while a marked broader peak with a maximum on 21 April occurs at the time of the Lyrids. This faithful reproduction of the meteor characteristics in the hourly rate of the transient echoes provided us with unmistakable evidence of the meteoric origin of at least a major proportion of these echoes.

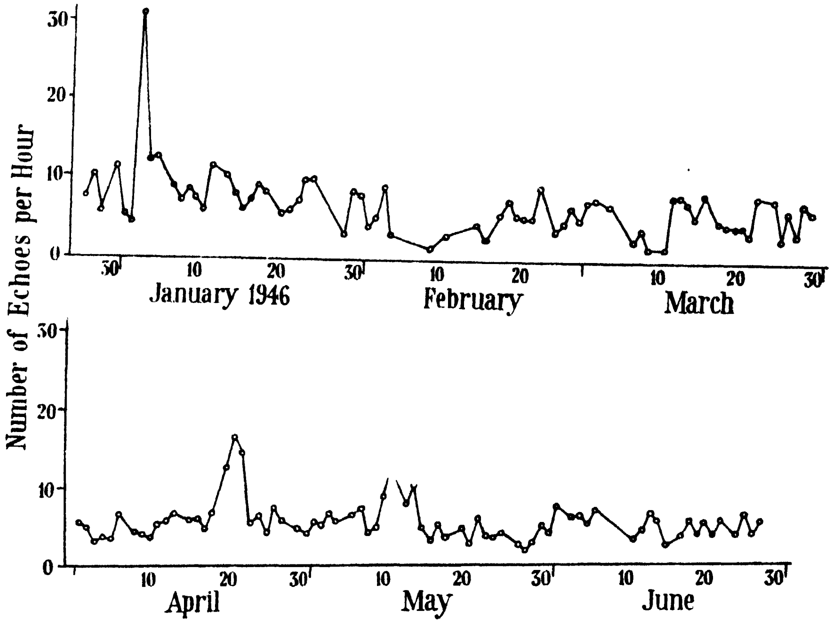


Figure 4. Mean hourly rate of occurrence of echoes. Vertical-beam station, Richmond Park.

(ii) ***Simultaneous echo observation and visual meteor watch,*** **20–22 *April* 1946.** A direct visual watch for meteors in the sky during the time of the Lyrid shower afforded further verification that the echoes are associated with meteors. These observations were made during 2030–2400 hrs. G.M.T. on each of the nights of 20–22 April. The sky was observed through a horizontal rectangular frame fixed vertically above the observer. The angular subtension of the frame roughly corresponded with the radar-beam coverage so that the field of view within the frame included most of the region in which a good radar response might be expected. It was possible to see a considerable region of the sky outside the rectangular frame although not to give it detailed attention. A celluloid sheet mounted in the frame and divided into zones enabled the apparent track of a meteor to be described approximately.

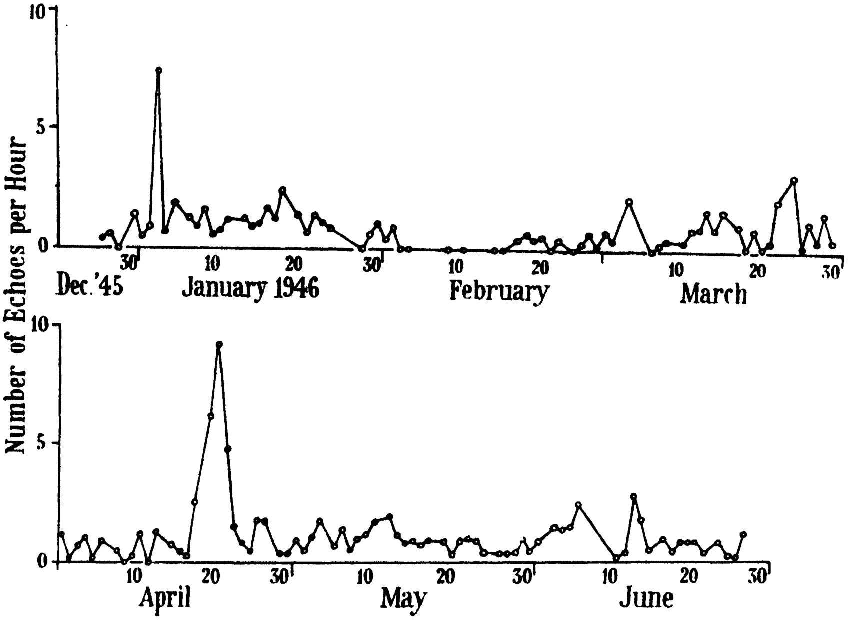


Figure 5. Mean hourly rate of occurrence of broad and multiple echoes. Vertical beam station, Richmond Park.

The sky was clear throughout the whole period of observation, with the exception of about twenty minutes on the second night when a filmy and broken patch of cirro-stratus entered the field of view, but this was not sufficiently dense to impair observation very seriously. The list of meteors, 13 in all, and particulars of the radar echoes, are given in table 2.

In table 2 there were 8 meteors whose tracks, produced if necessary, passed through the region of the viewing frame, and of these, 7 were associated with radar echoes. As to the remaining one, the observer stated that the track was extremely faint and he doubted whether he had genuinely seen a meteor. The visible tracks of three of the meteors which gave radar echoes were completely outside the frame, but in two cases the produced tracks passed through it.

We may conclude that certain meteors entering the radar coverage can definitely be associated with transient echoes. In addition to the radar echoes listed in the above table there were about seven times as many echoes with no meteors apparent in the sky. The existence of faint meteors which are observable telescopically although not visible to the unaided eye is well-known to astronomers. It is further possible that ionization trails undetected even by telescopes may originate radar echoes. The large number of echoes with no apparent visible meteor is therefore a result to be expected.

Table 2

METEORS				RADAR ECHOES				
No.	Date in April 1946	Time (G.M.T.)	Tracks passing through frame	Time (G.M.T.)	Dur. (sec.)	Range (km.)	Signal/Noise	Details of wide echoes
1	21	20.50.01	Yes	20.50.01	M	98	1½	—
2	21	20.59.20	Yes	20.59.20	3	121	Saturation	Beating 3 km. broad
3	21	21.57.00	No	No echo seen				
4	21	21.59.28	Yes	No echo seen				
5	21	22.27.51	Yes	22.27.53	M	112	1½	—
6	21	22.31.21	No	No echo seen				
7	21	22.32.39	Yes	22.32.40	1	113–116	1½	3 peaks moving outwards
8	21	23.39.43	Yes	23.39.51	M	98	1½	—
9	21	23.45.00	No	No echo seen				
10	22	22.08.03	Yes	22.08.05	½	107	2	2 km. broad
11	22	22.23.34	No	No echo seen				
12	22	22.40.27	No	22.40.27	M	102	5	2 km. broad
13	22	23.31.48	Yes	23.31.51	M	106	2	—

(M = echoes of duration less than 1 second.)

C. *The meteoric hypothesis*

The most important outcome of the investigations described above was the conclusion that at least a large proportion of the ionospheric echoes are caused by meteors. Skellett (1932, 1938) first suggested that the ionization caused by meteoric impact with the molecules in the upper atmosphere was the probable source of both the abnormal E ionization and the short duration echoes. Trowbridge (1907) had pointed out earlier the similarity between the visible radiation from meteor trails, which may persist many minutes, and the afterglow produced by electrical discharge in gases, which thus suggested that ionization had occurred. In their theoretical work, Lindemann and Dobson (1923), Sparrow (1926), and Maris (1929) all agreed that ionization would result from the impact between meteors and the molecules of the upper atmosphere. We shall now proceed to review our experimental data, formulating the ionospheric scatter echo characteristics more precisely and discussing in further detail the extent to which their properties conform to the meteoric explanation.

D. *Signal strength distribution*

We will first consider the distribution of signal strengths of the echoes. This is needed in taking the finite beam width into account to calculate the true heights from the apparent heights. Further, the distribution of equivalent echoing areas * is of interest for any determination of the density of ionization produced, although this subject is not pursued in the present paper.

The operator estimates the maximum amplitude reached by the echo pulse seen on the cathode-ray tube face. This maximum appears to be attained rapidly on first appearance. The receiver display is known to give approximate proportionality between echo height and the peak echo power. Figure 6 shows the observed distribution of echo powers in relation to the noise level. Owing to the liability of error in the operator's estimation of very low signal amplitudes,

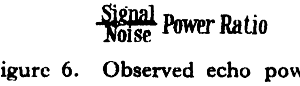


Figure 6. Observed echo power distribution, vertical-beam station A 2, June and July 1945. Note:—No experimental measurements above S/N=6.

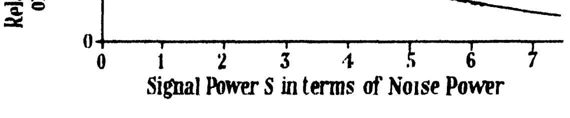


Figure 7. Comparison of computed and observed forms of distribution of echo powers ; assumed mean equivalent echoing area :—430 square metres.

---

* The equivalent echoing area A of an echo source may be defined by the relation :

$$\text{Echoed power flux at distance } r = \frac{A}{4\pi r^2} \times \text{incident power flux.}$$

we have considered only the cases where the ratio of signal power to mean noise power exceeds $1\frac{1}{2}$, and have therefore taken this as our lower limit of observation. The true distribution of echoing areas is modified, as regards the received power distribution, by the form of the radar beam, in which the sensitivity decreases progressively in regions further and further from the axis.

It was found that a reasonable agreement with the observed signal strength distribution could be obtained by assuming that the elementary probability δp that an echo will have a (maximum) equivalent echoing area between the limits A and $A+\delta A$ is of the form

$$\delta p = Ce^{-A/K} . \delta A,$$

where C and K are constants.

Since $\int_{A=0}^{A=\infty} dp = 1$, we derive at once that $C = 1/K$.

Further, the mean echoing area $\bar{A}$ is given by

$$\bar{A} = \int_{A=0}^{A=\infty} A . dp = \int_0 \frac{A}{K} . e^{-A/K} . dA = K,$$

whence we may write

$$\delta p = \frac{1}{\bar{A}} e^{-A/\bar{A}} . \delta A.$$

From the foregoing it may readily be shown that of a large number N of echoing sources, the number with echoing areas exceeding a chosen value A' would be

$$N . e^{-A'/\bar{A}} \quad \text{as} \quad N \to \infty.$$

The peak echo power received from a scattering source of equivalent echoing area A at a range R from the equipment is given by

$$S = \frac{P\lambda^2 G_T G_R}{64\pi^3 R^4} . A,$$

where P = peak power radiated, λ = wavelength, and G_T and G_R are respectively the power gains of the transmitting and receiving aerials in the direction of the scattering source.

Consider an elementary region of space of volume δv, such that the value of $G_T G_R/R^4$ remains sensibly constant throughout the region. Within this region, S is proportional to A, and the rate of appearance of echoes of power exceeding the lower limit S' will be proportional to

$$\delta v . e^{-A'/\bar{A}}$$
$$= \delta v . e^{-S'/\bar{S}},$$

where $\bar{S}$ is the mean signal power received from this region. For any given value of $\bar{A}$, and with a knowledge of the polar diagrams of the equipment and the transmitter power, we can derive the form of the variation of $S'/\bar{S}$ throughout space. By integration we may then determine the relative numbers of echoes exceeding given values of signal power.

We have performed the integration for the case of a station with a vertical beam, with the simplifying assumption that all the scattering sources occurred in a single thin layer at a height of 97 km. Comparing the theoretical results with the distribution observed at the appropriate station (during June 1945), we find that the value of $\bar{A}$ which gives closest agreement is in the neighbourhood of 430 m.^{2} The agreement between the observed distribution of echo powers and that computed by the above method is shown in figure 7.

It is to be emphasized that the choice of an exponential form of distribution is based purely on empirical grounds; also that the estimate of 430 m.^{2} as a mean equivalent echoing area only applies to a frequency of 73 Mc./s. ($\lambda = 4{\cdot}1$ m.). On account of the meteoric origin of the scattering sources, it is probable that the value of $\bar{A}$ will vary with the incidence of meteor streams of differing densities and velocities.

E. *Height distribution*

We have already seen in figure 3 that the range distribution of the echoes, as measured by the vertical stations, remained substantially the same for three separate periods during June and July 1945. Figure 8 shows the range distribution, analysed in 4 km. range bands, of all echoes observed in June and July 1945 by the vertical beam stations A 1 and A 2. In order to convert the smoothed distribution into a true height distribution, we must allow for the spread in range which results from the finite width of the radar beam. Suppose, for example, all echoes occurred at a fixed height of 100 km. Owing to the width of the radar beam, the distribution would have an abrupt near edge and then trail as shown in figure 9. The following method, illustrated in figure 10, was adopted to correct this effect. The observed range distribution was taken as a first assumed true height distribution, which we will call distribution "B". When the effect of the finite beam is applied we obtain a new distribution, "C". A distribution "A", determined by $AC = B^2$, was then deduced and used as the new assumed distribution for continuing the method by successive approximations. The solution obtained is shown in figure 11, the maximum being about 95 km., and 90% of all echoes being contained in a layer extending from 87·5 to 107·5 km. This corresponds well with the normal observed heights of meteors and their trains. Denning (1898) concluded that, in general, meteors appear at about 76 miles (approx. 122 km.) and disappear at about 51 miles (approx. 82 km.).

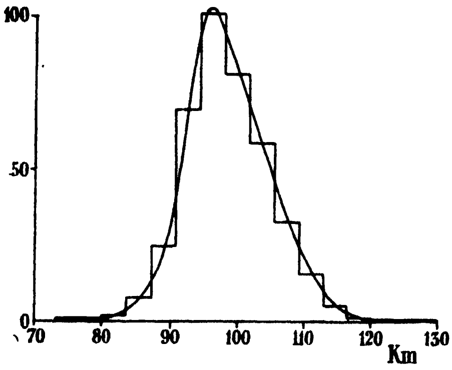


Figure 8. Mean range distribution of echoes for vertical-beam stations. June and July 1945.

Ordinate = relative number.
Abscissa = range.

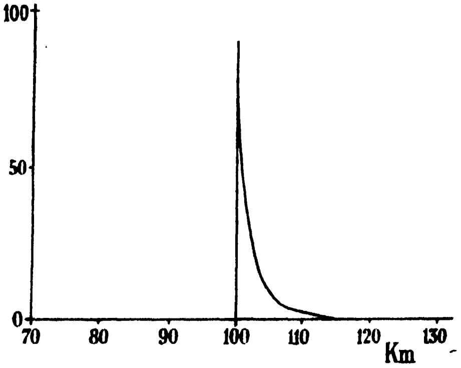


Figure 9. Range distribution of echoes from a thin layer at 100 km. height showing effect of radar beam width.

Ordinate = relative number.
Abscissa = range.

Trowbridge (1907) computed that the mean height of trains was 87 km., the limits for appearance and disappearance being 103 km. and 70 km. respectively. In more recent analyses of available data, Olivier (1942) deduced an average first height for night meteor trains as 102 km. and the end height as 74 km.; while Porter (1944) has given the mean heights for the beginning and end of the observed paths for sporadic meteors as approximately 103 km. and 86 km. respectively, although he also shows that there is a clear dependence of height on the elongation (the angular distance between the meteoric radiant and the apex of the earth's way) which determines the geocentric velocity.

It has been known from the earliest observations of the transient ionospheric echoes that they are roughly situated in the E region of the ionosphere. It is therefore of interest to consider to what extent the region of most frequent occurrence of echoes is related to the normal or abnormal E layers. Now, Appleton and Naismith (1940) found that the mean heights of the maximum ionization level for the normal E were 120 and 134 km. for summer and winter, while for the abnormal E they were 113 and 130 km. Since the region of maximum frequency of occurrence of the transient echoes is around 95 km. we conclude at once that it does not coincide in height with the maximum of either E layer. We may thus consider the region as a separate one and we refer to it as the *meteoric layer.* It will be seen that the shape of the height curve (figure 11) is similar to that of a Chapman region, which might be expected even for ionization produced by an external agency of this nature although the ionization is only of transient duration.

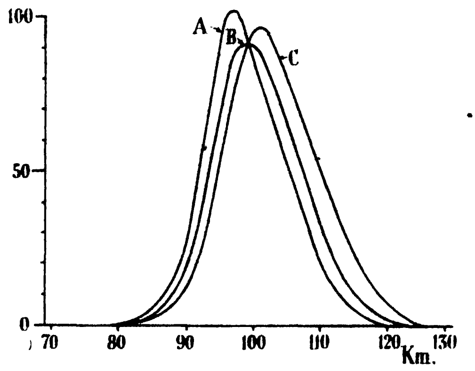


Figure 10. Illustrating method of deriving height distribution.
Ordinate = relative number.
Abscissa = range.

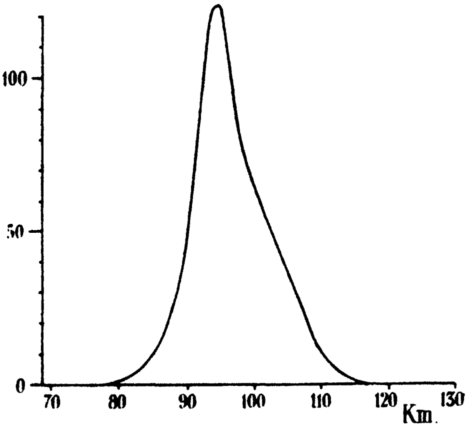


Figure 11. Mean height distribution of echoes for June and July 1945.
Ordinate = relative number.
Abscissa = range.

It is of interest to determine whether the range distributions are the same for transient ionospheric echoes of different characteristics in duration or complexity. In figure 12 the range distribution for momentary echoes (duration less than ½ sec.), is compared with that of echoes of longer duration, and that for echoes of complex appearance with that for single echoes. In both cases we see that longer duration echoes and complex ones show a range coverage extending to slightly greater heights. Variations in meteors, such as size, velocity and fragmentation, are assumed to be the controlling factors. Meteors of greater speed might be expected to produce ionization at greater heights where the rate of ionic recombination is slower due to reduced pressure and the echo consequently of longer duration. It does not appear justifiable to postulate origins of a different nature according to the echo characteristics, particularly as the differences in range distribution are small. Further support for this view is afforded by the determination of correlation coefficients for the entities in table 3 :—

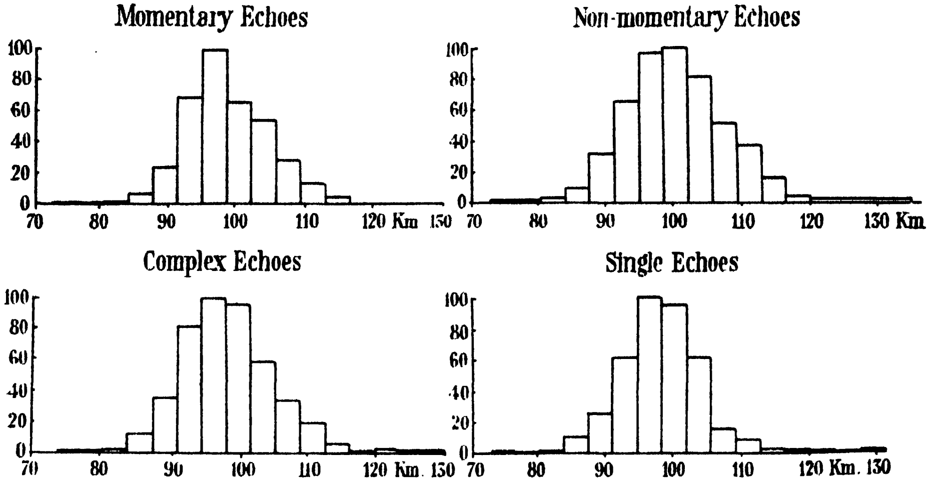


Figure 12. Range distributions of numbers of echoes of different characteristics. Vertical-beam station, June and July 1945. Ordinate = relative number. Abscissa = range.

Table 3

Entities		Period for analysis	Correlation coefficient
a Momentary echoes	*b* Longer duration echoes	30 days, by days	0·67
Single echoes	Complex echoes (multiple or broad)	30 days, by days	0·56
(No correlation was found between complex echoes and long duration echoes)			

Although the height distribution was essentially constant for three separate periods during June and July 1945, a small diurnal variation is discernible. The average range corresponding to the maximum frequency of occurrence of echoes has been plotted for each hour of the day in figure 13. This shows a diurnal variation of approximately 3 km., the maximum height being reached shortly after noon. The average rate of decrease of height which occurs from this time until after midnight is of the same order as that observed by Öpik (1937) from visual observation of night meteor trails, namely 0·3 km. per hour.

In order to investigate any seasonal variations of height we have plotted in figure 14 the mean range distribution for the vertical beam stations showing four successive periods between December 1945 and June 1946. There are some changes both in the width of the distribution and the height of maximum frequency of occurrence, in particular the period 1 February–30 March being characterized by a reduction in the tail of the distribution corresponding to echoes

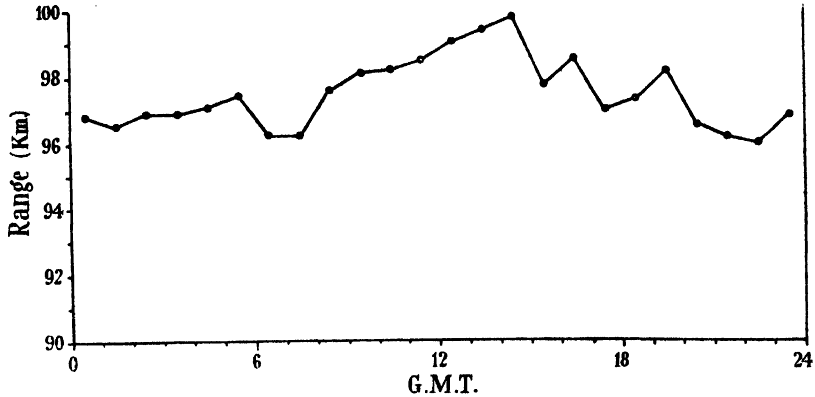
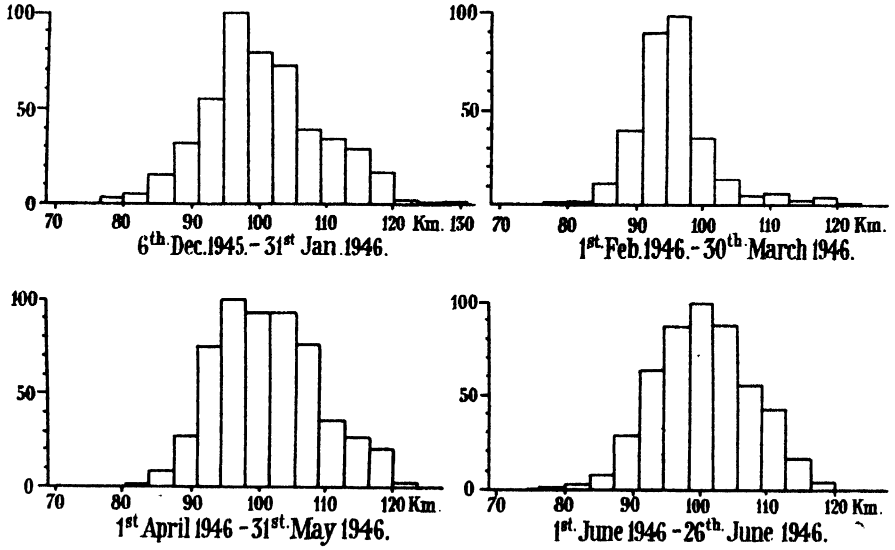


Figure 13. Diurnal variation in range of region of maximum density of echoes. Average of results for vertical-beam stations during June and July 1945.

at the greater heights. This shows an agreement with visual meteor observations by Öpik (1937), who stated that there is a seasonal variation in height with a minimum near 1 March. It should be noted that Porter (1944) has shown



6th Dec.1945.-31st Jan.1946.

1st Feb.1946.-30th March 1946.

1st April 1946 - 31st May 1946.

1st June 1946 - 26th June 1946.

Figure 14. Seasonal variations in range distribution of echoes for vertical-beam stations.

Ordinate = relative number. Abscissa = range.

that variations in meteor heights are not seasonal in the sense of being dependent on geophysical seasonal variations but are explicable in terms of the changes in average elongation which occur as the Earth moves along its annual orbit.

F. *Motion of scatter echoes*

The majority of the echoes appear stationary in range; approximately 2% of all echoes show a movement. Any change is most clearly detected in the case of the long duration echoes; 15% of echoes of duration exceeding 1 sec. show a movement, and the mean can be deduced with reasonable accuracy. These results are illustrated in figure 15, and the biggest group lies in the velocity interval between 0 and −5 km. per sec. The velocities are not, in general, sufficient for the movement to be explained by the velocity of the meteor itself. If the ionized column produced by the meteor is the source of the echo we should expect the maximum echo from the broadside-on view, and if the trail size remained constant and its position did not drift, the range would remain constant. One cause of the average reduction of range may well be the lateral expansion of the ionized column. Trowbridge (1907) noted that the average rate of diffusion appeared to be about 0·002 km. per sec. for periods of observation of the order of 10 minutes, but the rates were considered to be much greater immediately after formation of the trail. If the rates in the first few seconds are of the order of 1000 times greater this effect might account for the trend towards reduced range.

Drifts and distortion of the trail present another factor which can cause range movements. Visual observations by Olivier (1933, 1942) and others, have demonstrated that the winds in different strata show both different velocities and different directions. These velocities are generally of the order of 0·05 km. per sec., but the apparent rate of movement resulting from the shifting of the point of reflexion on the distorted train might be much greater. In this connexion we may refer to Eckersley and Farmer (1945) who considered the meteoric explanation unlikely from their observations that when a transient ionspheric echo is produced, a "mirror" type reflexion occurs, while after a few seconds there appears to be a number of widely spaced centres. This result, however, might well be expected from meteor trails, which are at first comparatively straight and then become distorted. The detailed description by Porter and Prentice (1939) of a long enduring trail may be taken as an example. The trail first appeared as a straight line of light but in a few seconds it was deformed and subsequently rotated through nearly 100°.

The observation of "whistles" in receivers tuned to carriers of short wave transmitters was made by Chamanlal and Venkataraman (1941), who reported some coincidence with visually observed meteors. These "whistles" were assumed to be due to the Doppler effect arising from the relative movement of the meteor itself. This apparent discrepancy with our pulse reflexion observations indicating that the majority of echoes seen show little or no range movement, and our consequent assumption that the echoes are broadside reflexions from the trails, requires explanation. We can suggest tentatively the following possibilities: (*a*) the echo from the head of the ionized column may be more readily detectable around 7 Mc./s., the radio frequency of the Indian observations, than at our radio frequency of about 70 Mc./s., and a continuous note presented aurally in any case provides a more sensitive means of detection than the visual presentation of a pulse on a cathode-ray tube; (*b*) the meteoric "whistles" may be, in fact, caused by the lateral expansion of the meteor trail; (*c*) that some of our observations of echoes moving in range (see figure 15) represent actual movement of the head of the ionized column, but that in general such echoes are missed in comparison with the stronger and more enduring reflexions from the column viewed from the side.

Chamanlal and Venkataraman found that E-layer returns were obtained by pulse reflexion methods on 7 Mc./s. immediately after the passage of a meteor as indicated by a whistle in a receiver tuned to a C.W. transmitter on a slightly different frequency. They considered this to be explained by the settling into a stratified layer of the ionization produced by the passage of the meteor. On a few occasions only they noticed a weak pulse-return with rapidly decreasing range occurring simultaneously with a strong Doppler whistle. In our view the most likely explanation is that the ionized air at the head of the meteor is generally missed by pulse reflexion methods because of the weak amplitude of the echo, the C.W. method with aural presentation of the Doppler note being more sensitive; further, the subsequent stationary pulse reflexion echo is from the meteor trail which cannot appear until the train has reached approximately the point of intersection of the normal from the observing station to the line of travel of the meteor.*

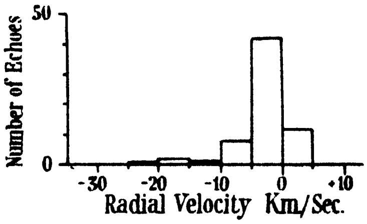


Figure 15. Distribution of radial velocity of moving echoes of duration exceeding ½ sec.
(Vertical beam station, July 1945.)

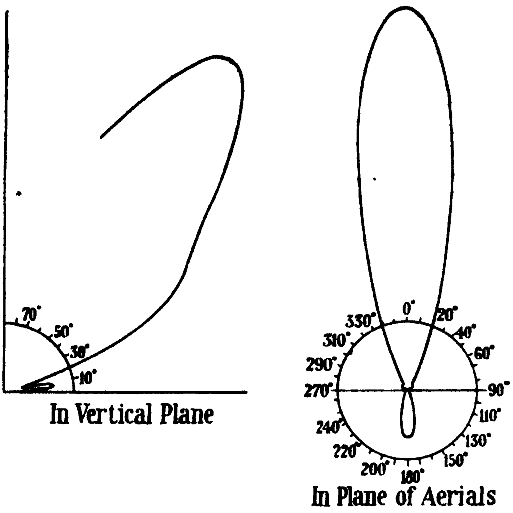


Figure 16. Polar diagram of radar signal power sensitivity.
(Inclined beam equipments, B_1, B_2, B_3.)

§4. INVESTIGATIONS WITH INCLINED RADIO BEAMS, JUNE–AUGUST 1945

In addition to the observations with vertical looking equipments, a number of experiments using equipments with inclined beams are described below. The equipments had single Yagis for the transmitting aerials and twin Yagis for reception. The combined polar diagram of radar signal power sensitivity is shown in figure 16, the maximum occurring at approximately 55° elevation.

---

* The above suggestion has since received confirmation in recordings with improved photographic techniques during the Giacobinid meteor shower of October 1946. These results are in course of publication in the *Monthly Notices of the Royal Astronomical Society*, and are outlined briefly in § 5.

A. *Observations during the Perseid shower*, 1945

An attempt to determine specific changes in rate of echo occurrence at an inclined beam station during one of the more notable meteor showers was made during the Perseid shower of August 1945. The equipment was sited in Richmond Park, and the watch was carried out daily from August 10–14 incl. between 0930 and 1230 hrs. G.M.T., the expected peak of the shower being 11–12 August. The set was maintained on a 90° bearing throughout. The Perseid radiant is R.A. 48°, Decl. +58°, and as observed from Richmond Park, should move in bearing from 300° to 320° and in elevation from 60° to 40° during the time of the watch. From the polar diagram of the aerial sensitivity shown in figure 16, and the range limits of observation, in this case from 80 km. to 200 km., it was deduced that the beam of the set covered directions at right angles to the radiant during the period of the watch. The authors are indebted to B. Rimmington for his assistance in carrying out the photographic recording and subsequent reading of the film.

One of the most striking results which emerged was the increased duration of the echoes, as compared with those of the previous months. The number of echoes of long duration rose to a pronounced peak on 12 August, when 24% of the echoes persisted more than 5 sec. This may be compared with the last week of July when less than 2% of the echoes were of more than 5 sec. duration. The frequency of occurrence of the echoes reached its maximum on 11 August. These two results are represented graphically in figure 17, and for comparison we have included those obtained during 26–31 July for the same daily period, set bearing, and recording method. The peak in the frequency of occurrence of echoes of more than 5 sec. duration is particularly marked. Although no direct visual observation of the meteor shower was attempted, it is known that the Perseids appear with no remarkable variations in numbers practically every August. The stream begins to cut the earth about the middle of July, and the maximum may be expected on 11 August. Meteors are still very frequent for the following two nights, after which there is a sharp decline (see, for example, Olivier, 1925). The observations therefore correspond well with the known astronomical characteristics.

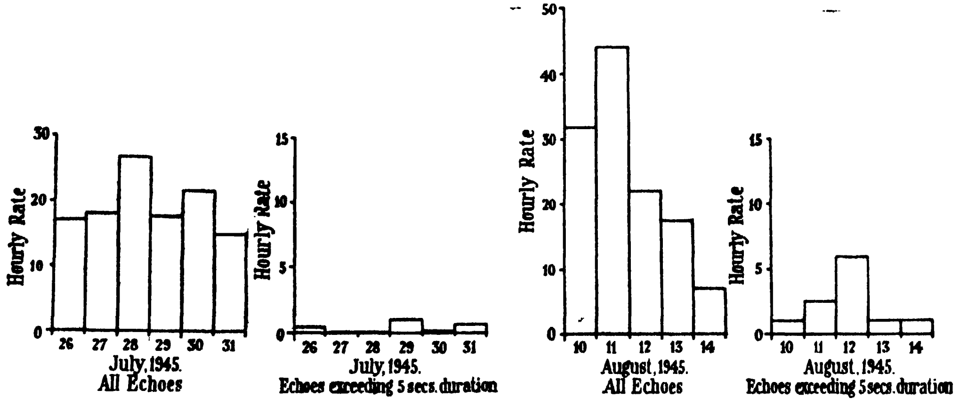


Figure 17. Hourly rate of occurrence of echoes.

B. *Observations with three inclined beam stations, June–July* 1945

Three similar beam equipments, which we shall refer to as B 1, B 2, and B 3, sited at Aldeburgh, Walmer, and Richmond respectively, were maintained on continuous watch throughout June and July 1945. As previously mentioned, this investigation was made possible through A.A. Command, who supplied the operators and administrative organization. The bearings of the three stations were initially chosen so that the axes of the radio beams intersected at a point about 100 km. in height and approximately equidistant from each station. The system is illustrated in figure 18. The purpose of this experiment was firstly, to fix the location in space of echoes seen simultaneously by the three stations by the range measurements, and secondly, to investigate any aspect effects, that is, variation of reflected signal according to the direction of incidence of the radio waves. It was realized that both of these aims might not be fulfilled, since if aspect proved a critical factor the chances of simultaneous observation of echoes by the three stations would be diminished. In the latter part of July the three stations were set on other bearings in order to compare the results for several directions, the axes of the three beams in these cases no longer having a common point of intersection. Visual recording was used throughout, with some additional checking of results by photographic recording. All the equipments were operated on 73 Mc./s. The experiments and the results obtained, for three periods in table 4, will now be discussed.

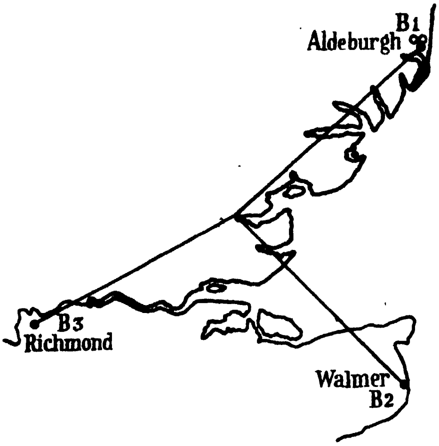


Figure 18. Equipments in operation, June and July 1945.

Table 4

Period	O.S. grid bearing of stations
1 June–17 July 1945	B 1, 230° ; B 2, 315° ; B 3, 62° (Beam axes intersecting)
17 July–26 July 1945	B 1, 180° ; B 2, 270° ; B 3, 90°
26 July–1 Aug. 1945	B 1, 180° ; B 2, 0° ; B 3, 90°

In comparison with the vertical beam stations it was found that the number of echoes had increased and the range distribution extended to greater ranges. These facts may be interpreted in terms of the greater area of the meteoric layer intercepted by the beams of the inclined beam equipments.

(i) *Aspect sensitivity of the echoes.* One of the most interesting features of the inclined-beam station results was presented by the marked diurnal variations in the frequency of occurrence of the echoes, the maxima and minima occurring at different times for the three stations respectively (see figures 19 and 21). This indicates at once that the sources of reflexion show marked aspect effects which are more apparent with inclined beams than with vertical. Confirmatory evidence of the aspect sensitivity of the sources of reflexion was provided by the rarity of simultaneous echo observations by the three stations. In order to allow for lags in operator's recording, a time difference of up to 5 sec. was permitted in regarding echoes entered in the logs of the respective stations as simultaneous.

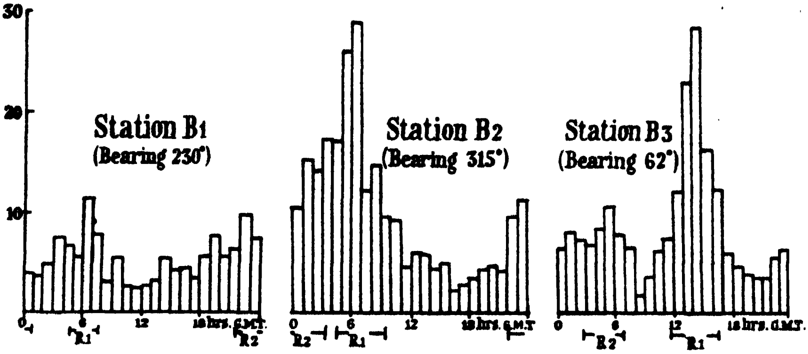


Figure 19. Diurnal variation of mean hourly rate of occurrence of echoes, 6 – 13 June 1945. Time at which radiants R_1 and R_2 are favourable are indicated by heavy lines.
Ordinate = mean hourly rate.
Abscissa = time G.M.T.

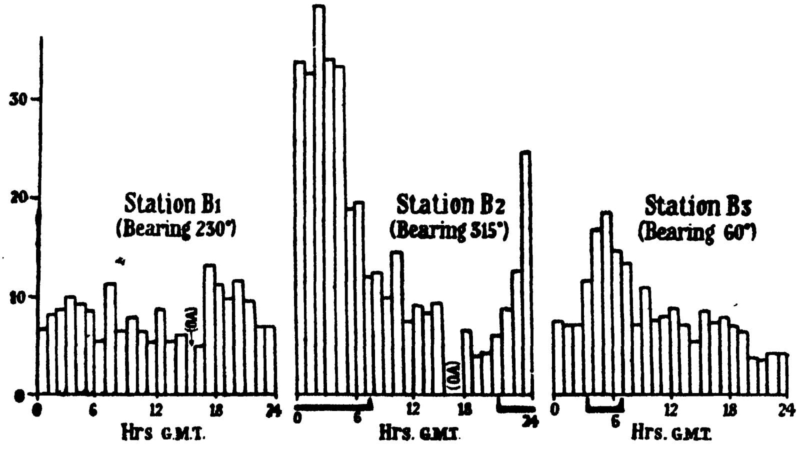


Figure 21. Diurnal variations of mean hourly rate of echoes, 26 July – 1 August 1945. Times at which the variant R is favourable are indicated by heavy lines.
Ordinates = relative number.
Abscissa = time G.M.T.
O.A. = out of action.

Now, in the period 1 June to 17 July 1945, on only four occasions was this criterion satisfied out of an average of over two thousand echoes observed by each station when they were all in operation at the same time. If the stations are considered in pairs, the average percentage of incidents observed by two stations, with not more than 5 sec. interval between recorded times, was approximately 3% of the total number of echoes seen by one station. In order to appreciate the theoretical chances of such coincidences, we must consider the extent of overlapping coverage. The probability of coincidence in different cases has been calculated assuming isotropic scattering sources in a single thin layer at 97 km. height, and an echo strength distribution as observed by the vertical beam stations (see § 3 D). The percentage coincidences calculated on this basis are compared below with the observed results for visual recordings.

Triple coincidences for B 1, B 2, B 3:—

Calculated probability = 3·1% of total number of echoes seen by all stations.

Observed coincidences = 0·003%.

Double coincidences for B 1 and B 3:—

Calculated probability = 10·5% of total number of echoes seen by both stations.

Observed coincidences = 1·6%.

Thus we see that even if we multiply the observed coincidences by a factor of 4 in the case of double coincidences and of 12 in the case of triple coincidences in order to allow for a 50% chance of an echo being missed by any one station, the calculated probability, based on the assumption of isotropic scattering sources, is greater than the observed probability. We may therefore conclude that the echoes show differential effects according to the direction of observation. This is further borne out by the coincidence percentage observed between B 1, the oblique beam station at Aldeburgh, and the vertical beam station A 1 (or A 2), situated on the same site. If the echoing sources are sensitive to the direction of incidence, then the probability of coincidence between B 1 and A 1 (or A 2) might be increased over that calculated on the basis of isotropic sources. This is actually the case as shown by the following figures.

Double coincidences for B 1, and A 1 (or A 2):—

Calculated probability = 1·2% of all the echoes seen by both stations.

(This figure is more liable to error than those quoted above owing to incomplete knowledge of the beam of the inclined station in the vertical direction.)

Observed coincidence = 3%.

This difference is accentuated by allowing for visual echoes missed. We conclude therefore that both the reduction of coincidences for stations on separated sites but with intersecting beams, and the increase for stations on the same site with partial overlap of coverage, as compared with the calculated values, accord with the assumption that the echo sources are sensitive to aspect.

(ii) *Meteoric explanation of aspect effects.* In terms of the meteoric explanation of the origin of the transient ionospheric echoes, the aspect effects are readily explicable. In discussing the range movements of the echoes we have already associated the main echoes with reflexions obtained by viewing the meteor trails along their normals. Pierce (1938, 1941) emphasized that meteor trails should, for radio waves, give specular reflexions appropriate to a reflecting cylinder. The maximum echoing area would therefore occur for a direction of incidence at right angles to the axis of the trail. We thus have at once the reason why the ranges of echoes generally remain nearly constant, namely that the ionized column is usually only sufficiently reflecting to be observable when viewed along the normal. The inability to obtain simultaneous observations of an echo by viewing the same region from different directions is at once explained. Further, the diurnal variations would be expected from the daily cycle of changes in bearing and elevation of the meteor radiants active at the time.

C. *Determination of meteor radiants*

As we pointed out above, the diurnal variations in frequency of occurrence of the echoes for the oblique stations according to their direction of look may be explained in terms of aspect sensitivity of the echoing source, the main echoes occurring when the trails are viewed along their normals. In this case it should be possible from the time of occurrence of the peaks in the diurnal variation to make some inference about the main directions of travel of the meteor streams and hence their radiants. Olivier (1925) has demonstrated that minor radiants exist in very great numbers scattered all over the visible heavens, and that any radiant is not worthy of consideration if it rests on observations of more than a few days. With these facts in mind we have chosen for analysis several periods of not more than one week, and have attempted to deduce only those radiants which might account for the main peaks in the diurnal variations.

We shall consider first the week 6–13 June 1945. The mean of the hourly rate for the three stations, B 1, B 2, and B 3, on bearings 230°, 315° and 62° respectively, are shown in figure 19. We note that the main peaks occur, for B 1 and B 2 at about 0630 hrs., and for B 3 at about 1430 hrs. In figure 20 we have drawn the coverage of possible radiant positions in the sky for each of these stations, at the time of their respective peak rates. It has been assumed that any possible radiants are at right angles to the radio beam. From a consideration of the beam characteristics and echo strength distribution the angular dimensions of the beam likely to contain 90% of the observed echoes were computed. The coverages shown in figure 20 represent possible directions at right angles to any part of this beam. We notice that the coverages for the peak hourly rate of

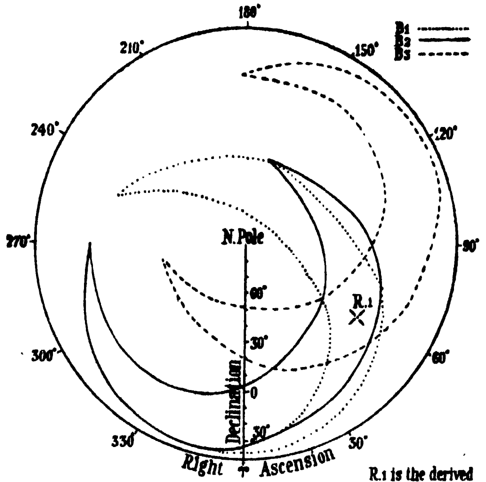


Figure 20. Coverage of possible radiant positions for main peaks in hourly rate of B 1, B 2, B 3.

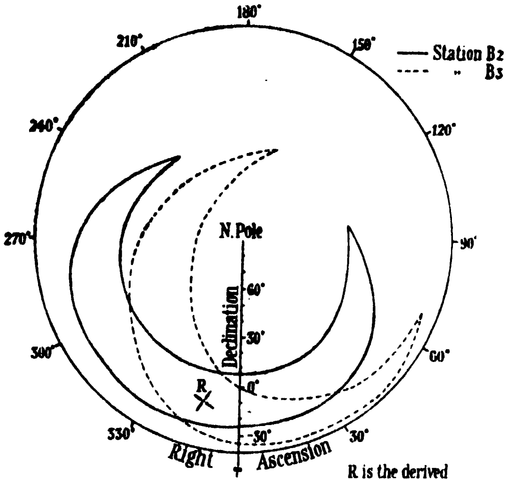


Figure 22. Coverage of possible radiant positions for main peaks in hourly rate of stations B 2 and B 3.

B 1, B 2, and B 3 all intersect in a common area, and we may assume this to contain the mean radiant, R 1. The position of this radiant appears to be within 10° of R.A. 58°, Decl. +5°. If we plot the secondary peaks in the hourly rate, namely at 2030 hrs. for B 1, and at 0530 hrs. for B 3, we can deduce by the same method a secondary radiant, R 2, within about 10° of R.A. 300°, Decl. −5°.

We must now verify that radiants in these positions give a satisfactory explanation of the observed hourly rate for both stations, and do not introduce any additional maxima to those which are observed. The times for which these radiants are within the station coverages may readily be determined graphically, the results being shown in table 5.

Table 5

Radiant	Site	Times for which radiant position is favourable (G.M.T.)
R 1 (R.A. 58° Decl. + 5°)	B 1	0440–0730
	B 2	0430–0930
	B 3	1140–1640
R 2 (R.A. 300° Decl. − 5°)	B 1	2130–0100
	B 2	2210–0310
	B 3	0240–0700

These periods have been marked below the time scale (abscissa) in figure 19 and they demonstrate that the radiants R 1 and R 2 give a satisfactory explanation of the times of the occurrence of the major and minor peaks for all three stations.

We shall next consider the diurnal variation of hourly rate for the period 26 July – 1 August, as shown in figure 21, and we shall here amplify the brief discussion of this case previously given by the authors (1946). The coverages of possible radiant positions corresponding to the marked peaks for B 2 at 0230 hrs., and B 3 at 0430 hrs., are shown in figure 22 and give a derived radiant in the neighbourhood of R.A. 345°, Decl. −10°. In figure 23 we have plotted the track of the radiant across the coverages, and marked the corresponding periods below the time scale for B 2 and B 3. The striking agreement is made even more remarkable by consideration of B 1 which exhibits a notably lower level of activity. As the coverage of this site never includes the radiant R in a favourable position, the absence of a marked peak in hourly rate is readily explained. The radiant R is plainly that of the δ Aquarids, a prominent stream of this epoch. The Perseid radiant may also be expected to have commenced activity towards the end of July. It can readily be shown that this radiant can only be viewed favourably by station B 1, between 1210–2320 hrs. G.M.T., and a small peak between these times is seen to occur.

In the above analysis we have not included a description of the effect of the radiants on the vertical beam station results. These stations give a ring type of coverage for possible radiant positions (see figure 24), and in many cases radiants may cross the coverage twice. This corresponds to the rising and setting of the radiant in the sky. This double intersection tends to even out the diurnal variation of hourly rate, which is in accordance with observation, the diurnal variations being of much less prominence than for inclined beam stations. In the case of only one of the radiants derived above is the radiant position so placed that it remains within the coverage for the whole of the time between the rising and setting of the radiant. This is the δ Aquarid radiant, and its track with respect to the coverage of a vertical beam station (A 1 or A 2) is shown in figure 24. We should therefore expect, in this case, the vertical beam station to give a clear peak between 2100 and 0630 hrs.; comparison with figure 25 demonstrates

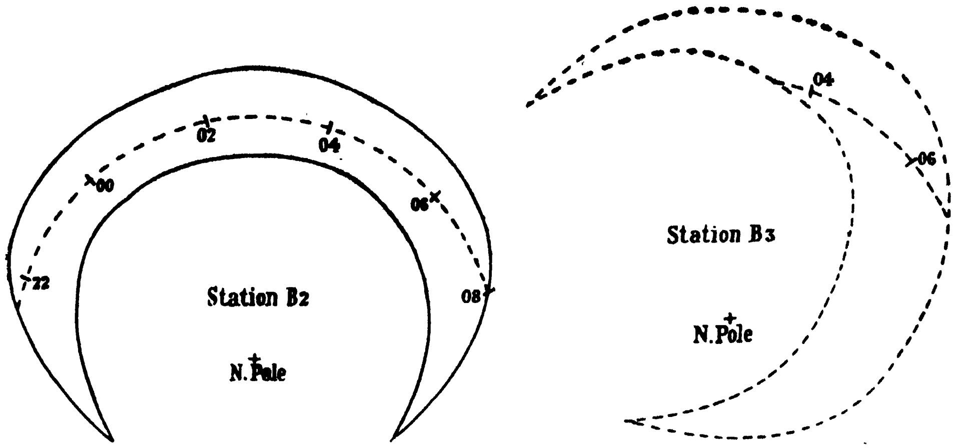


Figure 23. Tracks (dotted line) of meteor radiant R within coverages of stations B 2 and B 3. Time in hours G.M.T. marked on track.

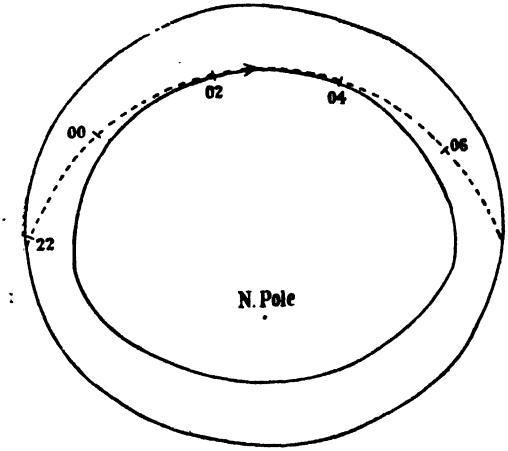


Figure 24. Track (dotted line) of meteor radiant within coverage of vertical beam stations A 1, A 1. Time in hours F.M.T. marked on track.

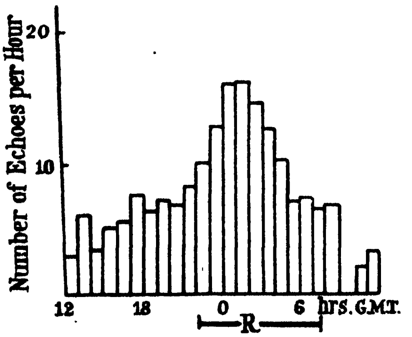


Figure 25. Diurnal variation of mean hourly rate of echoes for vertical beam station A 1. 26 July – 1 August 1945.
Times for which radiant R is favourable are indicated by heavy lines.

that this is the case, a more prominent peak being apparent for the vertical stations than at any other time in the June–July watch (see figure 2).

An accuracy in position of better than 10° cannot be claimed for the radiants determined above. Many other minor radiants must of course be active, but the width of the radio beams has not warranted any more detailed analysis at this stage. Further, there may be other factors which we are not in a position to assess as yet, such as the optimum angle of incidence of the meteor in the ionization layer for giving an effective radio response. Refinement of the method of observation by using narrower radio beams will undoubtedly lead to elucidation of these factors and to more accurate determination of radiants. An examination of the diurnal variations in rate of occurrence to determine whether there was any influence arising from the existing E-layer ionization yielded negative results. If such a variation exists, it is of a minor character at wavelengths of 5 metres in comparison with the major controlling factor, namely the direction of the radiant relative to that of the radio beam. This is not surprising since the wavelengths we have used are of the order of one-tenth the critical values for total reflexion from the E layers. As the critical wavelength is inversely proportional to the square of the electron density, we may infer that the electron densities giving rise to the transient echoes at 5 metres wavelength very greatly exceed those of the ionospheric E layer, and that existing ionization in this layer has little influence. A different result may well apply in transient echo observations at much longer wavelengths.

§ 5. THE GIACOBINID SHOWER OCTOBER 1946

An account of observations during this shower appears in a separate report by Hey, Parsons and Stewart,* and a brief outline of additional data and of certain points of importance to a general survey of the meteoric radar echoes will be given here. The rate of occurrence of the transient ionospheric echoes attained a remarkable peak between 0330 and 0400 hrs. G.M.T. on 10 October 1946, coinciding, as expected, with the maximum for visual observation of the shower as reported to us by J. P. M. Prentice. We found (§ 3, A) from observations during June 1945 that more echoes were observed at 55 Mc./s. ($\lambda = 5{\cdot}4$ m.) than on a similar equipment at 73 Mc./s. ($\lambda = 4{\cdot}1$ m.). The decrease in number of echoes observed as the radio frequency is increased had been noted by Appleton (1946). During the Giacobinid shower some observations were made at 212 Mc/s. ($\lambda = 1{\cdot}4$ m.) which is believed to be the highest frequency at which the meteoric echoes have been obtained. Between 0500 hrs. and 0545 hrs. G.M.T. on 10 October 1946, the number of echoes observed on the cathode-ray tube display of a 212 Mc./s. ($\lambda = 1{\cdot}4$ m.) equipment with vertically directed beam was seven as compared with about ten times that number on 64 Mc./s. ($\lambda = 4{\cdot}7$ m.) with a similar aerial system and power. A reduction in echoing area with increase of frequency is readily explicable in terms of gaseous ionization as the source of reflexion.

An important consequence of the introduction of improvements in resolution the of photographic recording method for the Giacobinid shower was the derivation of geocentric meteor velocities. On a 55 Mc./s. ($\lambda = 5{\cdot}4$ m.) equipment

---

* *Monthly Notices of the Royal Astronomical Society* (in publication).

wtih vertically directed beam, higher resolution was obtained by displaying a limited range band of 80 to 115 km. over the full width of the cathode-ray tube, and by using a 35-mm. film moving at 2·4 mm. per sec. instead of a 16-mm. film moving at 0·5 mm. per sec. Some of the recordings revealed a faint fast-moving echo prior to the main echo; an example is shown in plate 3, time 2·0–3·0 sec. This fits well with the views we expressed above (§ 3 F) in connexion with the Doppler whistles observed by Chamanlal and Venkataraman (1941), for we may associate the first faint trace with the ionization in the immediate vicinity of the approaching meteor. If we consider the meteor trail as a cylindrical column of ionized gas it is evident that the approaching head of the column, around the meteor itself, may originate a reflexion, although of smaller magnitude than the broadside echo from the column which occurs after the meteor has passed the point of minimum range. Similar weak echoes as the head of the meteor column recedes do not appear, presumably because whereas specular reflexion can occur from the rounded head of a cylindrical column as it approaches, only a diffracted echo of far smaller amplitude is possible as the head recedes. The secondary large echo shown in plate 3 may be due to distortion of the early part of the train resulting from drifts.

The faint tracks associated with the approaching meteor were mostly found to have range-time characteristics appropriate to uniform motion in a straight line, and the geocentric velocities could be determined from the formula

$$V=(R^2-R_0^2)^{1/2}/(T_0-T),$$

where V is the geocentric velocity, R the range at time T, and R_0 the minimum range at time T_0. The mean geocentric velocity derived from an analysis of 22 tracks was found to be 22·9 km./sec. which is in good agreement with the theoretically expected value of 23·7 km./sec. supplied to us by Dr. J. G. Porter. These results thus provided a further instance of the successful application of the radar methods to meteor observation.

§ 6. ACKNOWLEDGMENTS

We wish to acknowledge the valuable assistance placed at our disposal by the G.O.C. in C., A.A. Command, and we are indebted to Messrs. E. B. Britton, B. Rimmington and A. Sommi of the Operational Research Group, Ministry of Supply, for their help, and to all our assistants who took part in the subsequent analysis of results.

We wish to thank Sir Edward Appleton, F.R.S., and Mr. R. Naismith for supplying E-layer critical-frequency data and for helpful discussions, and to Mr. J. P. M. Prentice and Dr. J. G. Porter for their advice on meteor data.

Finally, we are indebted to the Chief Scientist, Ministry of Supply, and to the Superintendent of the Operational Research Group for their encouragement and interest in the investigations and for permission to publish this report.

REFERENCES

Appleton, E. V., 1945. *J. Instn. Elect. Engrs.*, **92**, I, 340.
Appleton, E. V., 1946. *J. Instn. Elect. Engrs.*, **93**, III A, 110.
Appleton E. V. and Naismith, R., 1940. *Proc. Phys. Soc.*, **52**, 402.
Appleton, E. V., Naismith, R. and Ingram, L. J., 1937. *Phil. Trans.*, A, **236**, 191.

Appleton, E. V. and Piddington, J. H., 1938. *Proc. Roy. Soc.*, A, **164**, 467.
Chamanlal and Venkataraman, 1941. *Electrotechnics*, **14**, 28.
Denning, W. F., 1898. *Nature, Lond.*, **57**, 540.
Eckersley, T. L., 1937. *Nature, Lond.*, **140**, 846.
Eckersley, T. L., 1940. *J. Instn. Elect. Engrs.*, **86**, 548.
Eckersley T. L. and Farmer, F. T., 1945. *Proc. Roy. Soc.*, A, **184**, 196.
Ferrell, O. P., 1946. *Phys. Rev.*, **69**, 32.
Herschel, A. S., 1911. *Observatory*, **438**, 291.
Hey, J. S. and Stewart, G. S., 1946. *Nature, Lond.*, **158**, 481.
Lindemann, F. A. and Dobson, G. M. B., 1923. *Proc. Roy. Soc.*, A, **102**, 411.
Maris, H. B., 1929. *Terr. Mag.*, **34**, 309.
Olivier, C. P., 1925. *Meteors.*
Olivier, C. P., 1933. *Proc. Amer. Phil. Soc.*, **72**, 215.
Olivier, C. P., 1942. *Proc. Amer. Phil. Soc.*, **85**, 93.
Öpik, E., 1937. *Harvard Annals*, **105**, 549.
Pierce, J. A., 1938. *Proc. Inst. Radio Engrs.*, **26**, 892.
Pierce, J. A., 1941. *Phys. Rev.*, **59**, 625.
Porter, J. G., 1944. *Mon. Not. R. Astr. Soc.*, **104**, 20.
Porter, J. G. and Prentice, J. P. M., 1939. *Brit. Astr. Assn. J.*, **49**, 337.
Schafer, J. P. and Goodall, W. M., 1932. *Proc. Inst. Radio Engrs.*, **20**, 1131 and 1941.
Skellett, A. M., 1932 *Proc. Inst. Radio Engrs.*, **20**, 1933.
Skellett, A. M., 1935. *Proc. Inst. Radio Engrs.*, **23**, 132.
Skellett, A. M., 1938. *Nature, Lond.*, **141**, 472.
Sparrow, C. M., 1926. *Astrophys. J.*, **63**, 90.
Trowbridge, C. C., 1907. *Astrophys. J.*, **26**, 114.

A STUDY OF THE NUCLEAR TRANSMUTATIONS OF LIGHT ELEMENTS BY THE PHOTOGRAPHIC METHOD

By C. M. G. LATTES, P. H. FOWLER and P. CUER,
H. H. Wills Physical Laboratory, University of Bristol

MS. received 2 January 1947

ABSTRACT. The photographic method has been employed in a study of the transmutation of lithium, beryllium, boron and oxygen by 900 kev. deuterons using the new Ilford Nuclear Research emulsion, type B1, for recording the tracks of the disintegration particles. The relation between the energy of a homogeneous group of particles and the mean range of the corresponding tracks in the emulsion has been determined both for α-particles and for protons. With protons, the mean stopping power of the emulsion, relative to standard air, varies from about 1800 at 2 Mev. to 2000 at 13 Mev. Curves showing the stopping power of the emulsion as a function of range are given for both protons and α-particles.

The experiments show that the uncertainty in the energy of an individual proton, as deduced from the length of the track which it produces in the emulsion, is only slightly greater than that due to straggling. In view of the very refined geometry which can be employed when using the photographic method, it follows that it can be applied in experiments of the highest precision.

The observations with a beryllium target give evidence for the existence of a group of particles, not previously observed, which we attribute to protons from the reaction ${}^{9}_{4}\mathrm{Be}+{}^{2}_{1}\mathrm{H}\rightarrow{}^{*10}_{\ 4}\mathrm{Be}+{}^{1}_{1}\mathrm{H}$, in which the ${}^{10}_{4}\mathrm{Be}$ nucleus is formed in an excited state of energy $3{\cdot}40\pm0{\cdot}08$ Mev.

The difference in the appearance of the tracks of protons and α-particles in the B1 emulsion is sufficiently great to enable the two types to be distinguished by inspection. It is thus possible to make measurements on groups of protons in the presence of α-particles of the same range.

§ 1. INTRODUCTION

In a recent publication (Powell, Occhialini, Livesey and Chilton, 1946), which we shall refer to as (I), an account has been given of the prototype of new photographic emulsions for recording the tracks of fast charged particles from nuclear transformations. These emulsions are now being produced in the research laboratories of Ilford Ltd. and are described as *Nuclear Research Emulsions* (N.R.). The visibility of the tracks in the N.R. emulsions is greatly superior to that obtained with ordinary half-tone plates, and the previous work shows that the field of application of the photographic method is thus extended and its precision improved.

In this paper we describe experiments in which we have investigated the characteristics of the emulsion in more detail, giving particular attention to the following technical aspects of the method:—

(*a*) The range-energy relation for protons and α-particles in the emulsion.

(*b*) The energy "resolving power ' of the method.

(*c*) The "discriminating power" of the emulsion.

In the case of the ordinary half-tone emulsion it has been shown (Powell, 1943; Guggenheimer, Heitler and Powell, 1946), that the length of the track of a proton ot an α-particle is proportional to its range in standard air for particles of all energies in the interval from 1 to 13 Mev. Owing to the much higher concentration of silver halide relative to gelatine, this simple relationship is not valid for the N.R. emulsions. It is therefore important to make experiments to determine the range-energy relation for the different particles so that the distribution in length of the tracks measured in any experiment may be transformed to a distribution in energy.

For this purpose we have made measurements with various homogeneous groups of particles of known energy. In addition to the α-particles from the naturally occurring radio-active nuclei, we have employed the groups of protons and α-particles emitted from the light elements under bombardment with 900 kev. deuterons. Values of the energy release, Q, in these reactions are in many cases now established with high precision. It follows that the energy of the corresponding particles can be calculated from the known energy of the primary deuterons and from the direction of emission relative to the primary beam at which the observations are made. By comparing the mean lengths of the tracks, produced by homogeneous gr ups of protons and α-particles, with the corresponding values of the energy, we can deduce the relation between the energy and the range in the emulsion for particles of the different types.

Having established the range-energy relations, we can proceed to examine the energy "resolving power" which the method allows. For this purpose we transform the observed range distributions to an energy scale. The width of the resulting peaks at half maximum can, by analogy with the corresponding optical problem, be taken as a measure of the capacity of the method to distinguish between homogeneous groups of similar particles of different energy, and in conformity with previous workers we define this as the *energy resolving power*.

The third aspect of the method, the discriminating power, refers to the possibility of distinguishing different types of particles of the same range by the differences in the numbers of grains per unit length in the corresponding tracks. The performance of an emulsion in this respect depends, amongst other factors, on the mean grain size of the silver halide particles and varies from one type of emulsion to another. Our present experiments are confined to observations with the B1 emulsion, in which the mean grain size is $0{\cdot}4\,\mu$, and we show that it is possible to distinguish protons and α-particles with high efficiency. Thus, if we examine a plate containing a group made up of α-particles and protons of the same range, we can decide in 90% of the cases, by simple inspection, whether a track is due to a proton or an α-particle. In the remaining 10% of the cases the correct allocation is uncertain. For many purposes this degree of discrimination is sufficient, but further improvement is desirable since, with the present emulsions, there may be an ambiguity in the interpretation of individual events such as occur in the nuclear explosions produced by exposures to cosmic radiation or high energy γ-rays.

This is achieved in the C1 and E1 types of emulsion, where the difference between the tracks of protons and α-particles of the same range is greatly accentuated. These emulsions have the disadvantage, however, that for work where exposures of saveral months are employed there is a marked fading of the latent image. The tracks produced by lightly ionizing particles tend to disappear first, and in certain experiments this fact can be employed to increase the discrimination between, for example, tritons and α-particles. With E1 emulsions the tracks of protons disappear almost completely after one week and, with C1, after two weeks.

§ 2. EXPERIMENTAL METHOD

The apparatus employed is similar to that described in (I), with small modifications shown in figure 1. The primary beam of fast deuterons from the Cambridge high-tension generator, after magnetic analysis to remove molecular ions, falls on a molybdenum plate in which there is a central aperture 3 mm. in diameter. The defined beam strikes the layer of target material carried on a shaped metal block which can be rotated on a cone. The different surfaces, carrying thin layers of various target materials, can thus be rotated into the beam in succession.

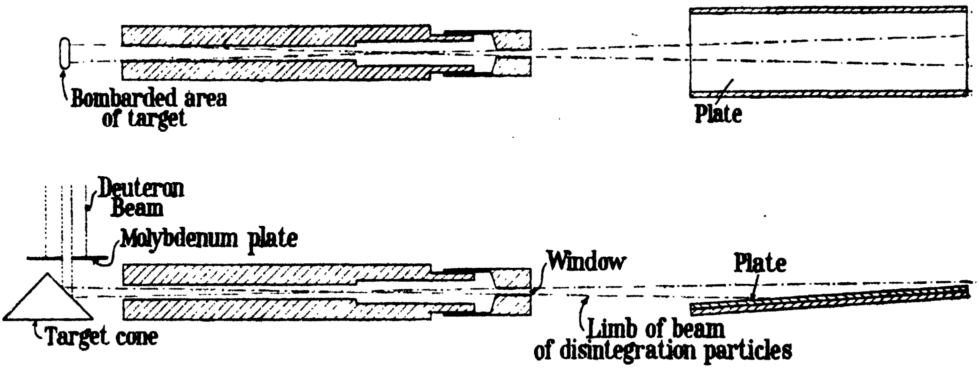


Figure 1.

In making exposures, targets of the following light elements were employed:—

(a) *Beryllium.** A thin layer of metallic beryllium was deposited on one face of the brass block by evaporating the metal from a tantalum boat *in vacuo*. This target had a brilliant metallic lustre and its stopping power was of the order of 0·5 mm. of air. The plates obtained therefore record disintegration particles produced by a beam of deuterons nearly homogeneous in velocity, the loss of energy of the particles in passing through the beryllium layer being less than 20 kev. for protons and 80 kev. for α-particles.

(b) *Lithium oxide.* This target was produced by burning lithium in air and allowing the white smoke of lithium oxide to condense on the target surface. Thick targets of this material were employed since it was found to be rapidly removed by the impact of the beam. In order to avoid contamination of the target chamber by lithium, this material was employed last in the series of exposures and was mounted on a separate supporting cone.

(c) *Boron.* A layer of amorphous boron was applied to the copper surface by grinding the material in a mortar with xylol and painting the resulting material on to the copper with a brush.

The disintegration particles were observed in a direction making an angle of $90 \pm 1°$ with the primary stream of deuterons. Simple calculations show that in all cases the conditions of observation are well within the limits of "good geometry" as defined by Livingston and Bethe (1937). The particles emerge from the vacuum chamber through a window of formvar having a stopping power of 3·1 mm. of air for polonium α-particles. The pressure in the camera containing the plate is reduced, during the exposures, to such a value that the loss of energy suffered by the particles in passing between the window and the plate is negligible.

The formvar window, in spite of its low stopping power, was sufficiently strong to withstand a pressure difference of more than 1 atmosphere. This result was achieved by covering the film with a thin protective layer of silver deposited by evaporation *in vacuo*. Ordinary windows of the formvar type are subject to deterioration through oxidation processes promoted by ionization. In addition to giving a marked increase in strength, the silver therefore serves also to protect the window from this type of chemical attack. The presence of the window allows the removal of exposed plates from the "camera" and the reloading of new ones without admitting air to the vacuum system containing the targets. The speed at which exposures can be made is thus greatly increased.

The angle of approach at which the disintegration particles, emerging through the window, meet the photographic emulsion can be varied at will, and we have found a suitable value to be 4°. Such a disposition ensures, firstly, that the proportion of particles entering the plate which are deflected from their original direction, through small-angle Coulomb scattering, to such an extent that they return to the surface of the emulsion is very small. If a particle leaves the emulsion its range cannot be determined, and such particles are not measured. Secondly, a small angle of approach ensures that relatively few of the fast particles traverse the whole thickness of the emulsion to enter the glass.

The thickness of the emulsions employed in the present experiment was $40\,\mu$. There are no particular difficulties in processing such plates, but the following

* We are indebted to Dr. J. W. Mitchell for the preparation of the beryllium target.

procedure, for which we are indebted to Dr. G. P. S. Occhialini, is adopted as standard practice and is of importance in dealing with thicker emulsions of the order of 100 μ. The plates are developed in Ilford I.D.19 (x-ray developer), at a dilution of 1 of developer to 3 of water, for 35 minutes at 18° C., with constant rocking. For this purpose a mechanical device is employed which gives regular sudden changes of tilt to the dish and provides laminar flow of the developer over the plate. After development, the plate is washed for 1 min., bathed for 1 min. in 2% acetic acid, and fixed. A novel feature of the fixing, which presents the most serious technical difficulties with these emulsions, is that it is important to rock the fixing bath. The emulsions contain roughly a hundred times as much silver per unit area as an ordinary plate, and it becomes important to ensure that sufficient fixing salt is provided. We find it an advantage to employ freshly prepared fixing solution of the following composition: two parts of saturated sodium thiosulphate solution and one part of Kodak acid fixing bath at standard concentration added to six parts of water.

The bath is renewed several times during fixation, the plate being washed for a minute whilst the solution is being changed. It is of some advantage to maintain the surface of the emulsion face downwards in the bath to facilitate the removal of the heavy silver salts from the emulsion, and for this purpose the plate is supported by attaching a suction pad, carried by a metal stem, to the glass.

Considerable variation of development time can be tolerated without appreciable changes in the visual quality of the image seen under the microscope. As the development time increases, the background grains become numerous, but the tracks are denser. In our experience it is better for the development time to be too long than too short, especially for work with protons, when a sharply defined beginning to the tracks is important. A certain concentration of background grains is useful in allowing the surfaces of the emulsion to be easily distinguished, since this allows a quick decision as to whether or not a particular track has passed out of the emulsion. If good discrimination between different particles is important, it is of some advantage to under-develop.

§ 3. EXPOSURES AND METHOD OF MEASUREMENT

The following are the particulars of the exposures given to the plates on which the present measurements were made:—

No.	Target	Stopping power in air equivalent (mm.)	Beam current (μa.)	Time of exposure (min.)
1	Beryllium	0·5	50	60
2	Lithium oxide	Variable	70	10
3	Boron	Thick	70	16

The plates were examined using a variety of objectives and eye-pieces, different observers making observations on the same plate with different magnifications. We draw attention to the fact that, when using binocular microscopes, the overall magnification depends on the interocular distance employed and, therefore, varies with different observers. Individual calibration of the micrometer scale for each observer is therefore essential. We take as our standard of length a 1-mm. graticule by Cooke divided into one hundred parts.

In determining the length of the tracks we chose to measure the length of the projection of the trajectory of the particle on the plane of the surface of the emulsion. The value so obtained is then multiplied by $\sec\theta$, where θ is the average angle of approach of the particles to the emulsion. The errors introduced by ignoring the effect of changes in the angle of "dip" of the tracks are very small. It would be possible to make appropriate corrections but, owing to the frequent small changes in the direction of the particles, due to Coulomb scattering, the work of measurement would be very greatly complicated. We therefore preferred to sacrifice a small degree of precision in favour of the very important advantage of speed of measurement. In making the measurements we define as the *length* of a track the distance between the extremities of the first and last grains recognized as belonging to it, and we reorientate the scale to lie parallel to the projection of the track whenever changes in direction occur. We thus take account of the changes in direction of the projection of the track, although the changes in "dip" are ignored. The average error in the range measurements due to this approximation is much less than the fluctuations due to straggling.

§ 4. EXPERIMENTS WITH LITHIUM OXIDE

A typical photo-micrograph of a small area of the surface taken with the target of lithium oxide is shown in plate 1. The prominent dense tracks are due to α-particles and the thinner tracks produced by protons can be clearly distinguished from them. The finite depth of focus of the objective prevents the tracks from being sharply in focus throughout their length, but the ends can in many cases be distinguished. We can therefore identify the reactions leading to the emission of particular particles.

In figure 2*a*, we show the distribution in length of 2000 tracks taken from measurements on this plate. Three different magnifications were employed. The short tracks in the range from 10 μ to 50 μ were measured with a Cooke × 95 O.I. achromatic objective with × 15 eye-piece, in which one small division in the eye-piece corresponds to 0·442 μ.* The tracks from 50 μ to 100 μ were measured with the same objective and a × 10 eye-piece (1 div. = 0·555 μ), and the longest tracks with a Leitz × 65 apochromat with × 6 eye-piece (1 div. = 0·984 μ). The resulting measurements are transformed to length in microns and plotted in the same diagram, but the relative numbers of particles in the three ranges have not been adjusted to correspond to the true values. The main features in the observed distribution are the succession of groups, marked (i) to (viii) inclusive, which we attribute to the following well established reactions:—

(i) corresponds to the continuous distribution of α-particles from the reaction ${}^7_3\mathrm{Li}+{}^2_1\mathrm{H}\rightarrow 2\,{}^4_2\mathrm{He}+{}^1_0\mathrm{n}$. Near the long-range end of this continuous distribution there is a clearly defined peak (ii) which we attribute to the alternative reaction ${}^7_3\mathrm{Li}+{}^2_1\mathrm{H}\rightarrow{}^5_4\mathrm{He}+{}^4_2\mathrm{He}$ corresponding to the formation of unstable ${}^5_2\mathrm{He}$ nuclei and their subsequent disintegration into α-particles and neutrons (Williams, Haxby and

---

* For a particular interocular distance. The difference between the observed and the nominal magnification is due to the use of the binocular eye-piece attachment.

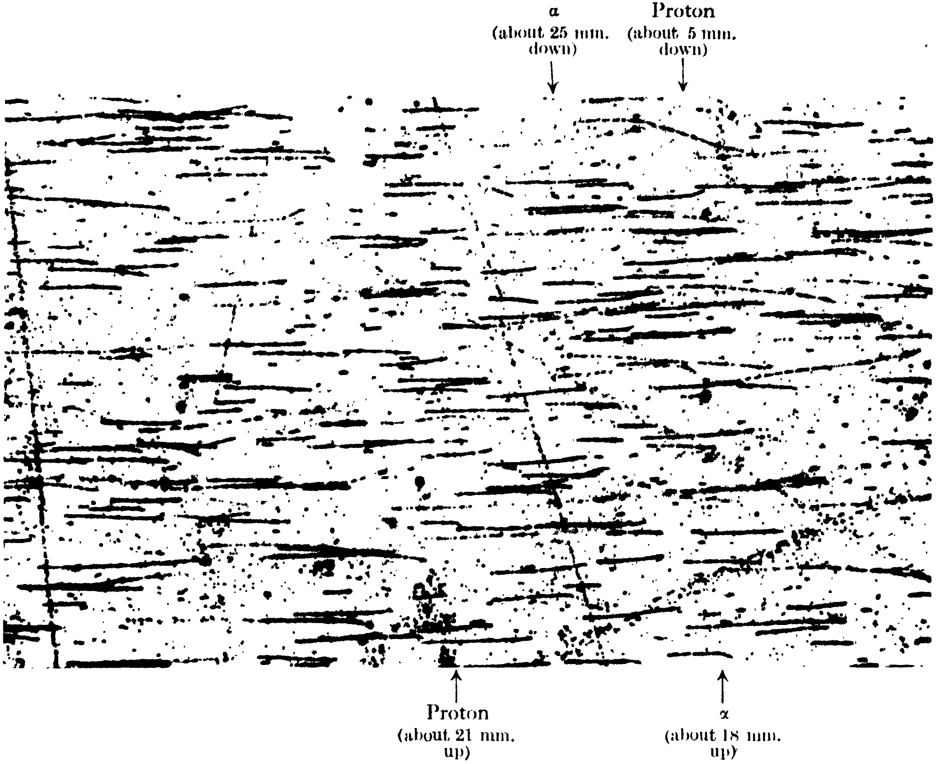


Plate 1.

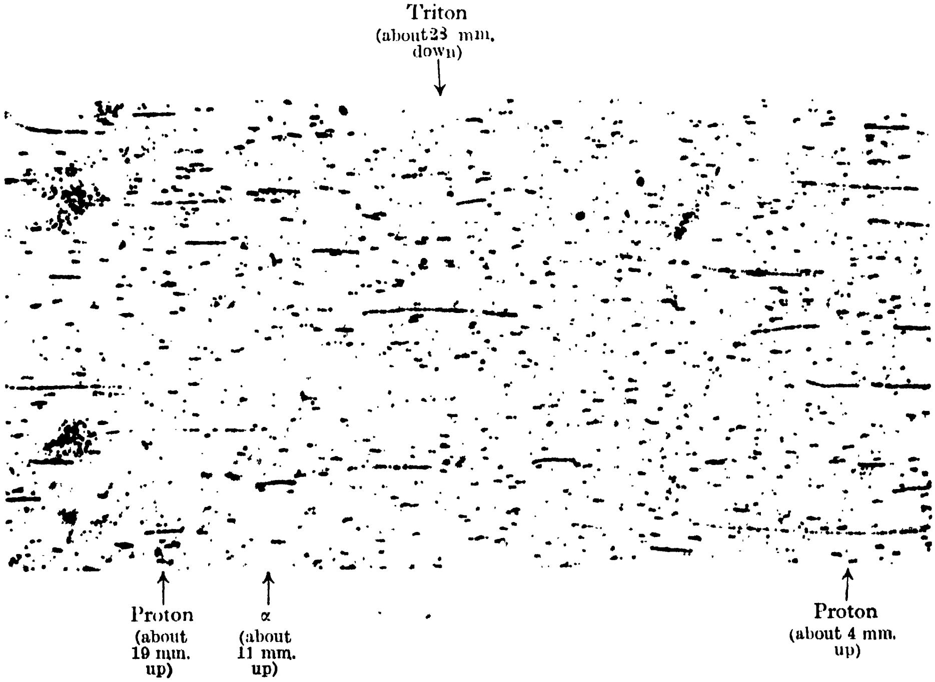
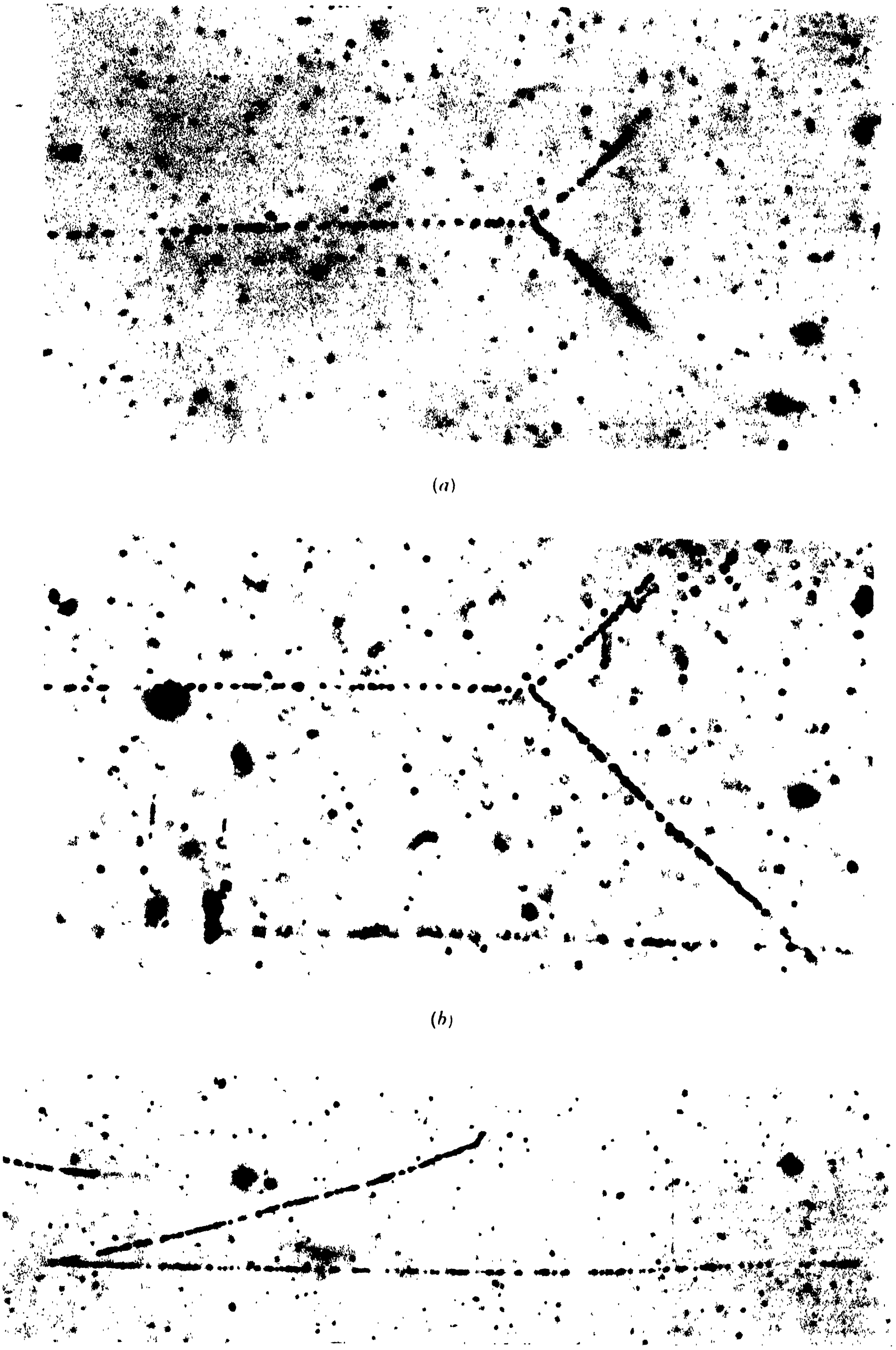


Plate 2.



(a)

(b)

(c)

Plate 3.

Shepherd, 1937 a). From the range-energy relation for α-particles, established in the present experiments, we can deduce the energy of the α-particles produced in the reaction and hence the corresponding energy release, Q. We thus obtain the value $Q = 13{\cdot}43$ Mev., which corresponds to a mass of 5·0146 m.u. for ${}^5_2\mathrm{He}$. The corresponding values deduced from the experiments of Williams *et al.* (1937 b) are $Q = 12{\cdot}66$ Mev., ${}^5_2\mathrm{He} = 5{\cdot}01543$ m.u. According to the present determination of its mass, a ${}^5_2\mathrm{He}$ nucleus should dissociate into an α-particle and a neutron with a release of energy of 1·3 Mev. In order to confirm the existence of a group of α-particles corresponding to the formation of ${}^5\mathrm{He}$, we made a separate exposure with a thin target of lithium oxide, and the results of measurements on the plates so obtained are shown in figure 2(*d*). It will be seen that there is an improved resolution of the α-particle group. The difference between the distribution at the lower ranges is due to the fact that in the second series of measurements the tracks of protons and α-particles were accepted without distinction, whereas the first observations were confined to α-particles.

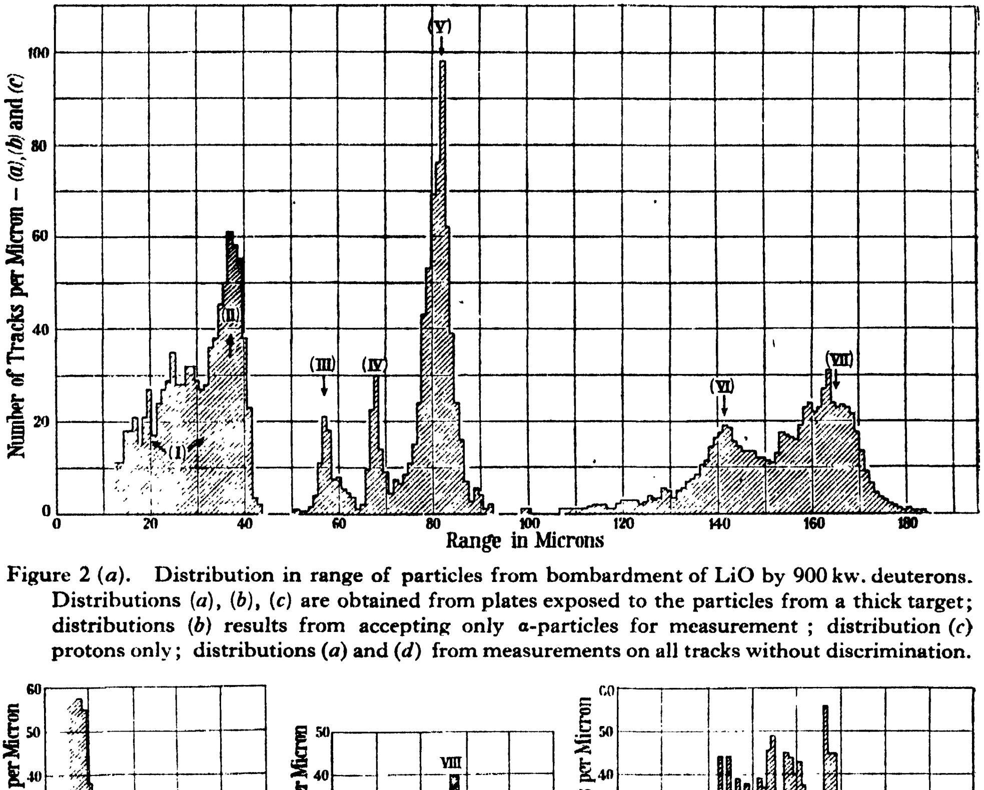


Figure 2 (*a*). Distribution in range of particles from bombardment of LiO by 900 kw. deuterons. Distributions (*a*), (*b*), (*c*) are obtained from plates exposed to the particles from a thick target; distributions (*b*) results from accepting only α-particles for measurement ; distribution (*c*) protons only ; distributions (*a*) and (*d*) from measurements on all tracks without discrimination.

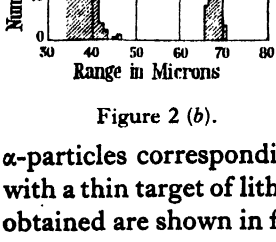


Figure 2 (*b*).

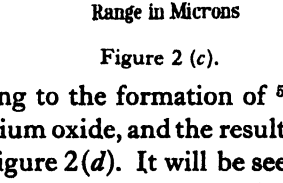


Figure 2 (*c*).

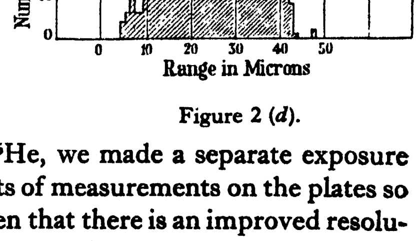


Figure 2 (*d*).

We attribute peak (iii), figure 2(*a*), to protons from the reaction $^{16}_{8}O + ^{2}_{1}H \rightarrow ^{17}_{8}O + ^{1}_{1}H$, the $^{17}_{8}O$ nuclei being left in the ground state. Peak (iv) is due to α-particles from the reaction $^{6}_{3}Li + ^{2}_{1}H \rightarrow ^{4}_{2}He + ^{4}_{2}He$. The peak (v) may be due to protons either from the reaction $^{2}_{1}H + ^{2}_{1}H \rightarrow ^{3}_{1}H + ^{1}_{1}H$, or from $^{12}_{6}C + ^{2}_{1}H \rightarrow ^{13}_{6}C + ^{1}_{1}H$, the expected difference in energy of the protons in the two cases being only 30 kev. in the conditions of our experiments. We think it probable that it is due mainly to the presence of carbon because of the method of preparing the target and the relatively great intensity of the group as compared with that of nearly the same range obtained in the experiments with beryllium. Finally, peaks (vi) and (vii) are due to protons from the reaction $^{6}_{3}Li + ^{2}_{1}H \rightarrow ^{7}_{3}Li + ^{1}_{1}H$, the $^{7}_{3}Li$ nuclei being produced in the ground state or in the well known excited state at 440 kev.

Table 1

Reaction	Q (Mev.)	E (Mev.)	R (microns)	d (microns)	Ref.
(i) $^{7}_{3}Li + ^{2}_{1}H \rightarrow 2\,^{4}_{2}He + ^{1}_{0}n$					(1)
(ii) $^{7}_{3}Li + ^{2}_{1}H \rightarrow ^{4}_{2}He + ^{5}_{2}He$	12·7 (13·43)	7·3 (7·76)	39·5	0·10	(5)
(iii) $^{16}_{8}O + ^{2}_{1}H \rightarrow ^{17}_{8}O + ^{1}_{1}H$	1·95±0·06	2·59±0·06	59·6	0·3	(2)
(iv) $^{6}_{3}Li + ^{2}_{1}H \rightarrow ^{4}_{2}He + {}_{2}He$	22·20±0·04	11·32±0·02	70·8	0·2	(3)
(v) $^{2}_{1}H + ^{2}_{1}H \rightarrow ^{3}_{1}H + ^{1}_{1}H$ $^{12}_{6}C + ^{2}_{1}H \rightarrow ^{13}_{6}C + ^{1}_{1}H$	3·98±0·02 2·71±0·05	3·20±0·02 3·23±0·05	83·9	0·1	(4) (2)
(vi) $^{6}_{3}Li + ^{2}_{1}H \rightarrow {}^{*}{}^{7}_{3}Li + ^{1}_{1}H$	4·58±0·12	4·51±0·12	147·0	0·2	(1)
(vii) $^{6}_{3}Li + ^{2}_{1}H \rightarrow ^{7}_{3}Li + ^{1}_{1}H$	5·35±0·12	4·96±0·12	170·0	0·2	(1)
(viii) $^{16}_{8}O + ^{2}_{1}H \rightarrow {}^{\dagger}{}^{17}_{8}O + ^{1}_{1}H$	(1·12)	(1·81)	27·9		(2)

(1) Rumbaugh, Roberts and Hafstad (1938).
(2) Cockcroft and Lewis (1936 b).
(3) Smith (1939).
(4) Oliphant, Kempton and Rutherford (1935 a).
(5) Williams, Haxby and Shepherd (1937 a).

The bracketed values are those deduced from the present experiments.

A correction for the "thick target effect" has been applied to the results for every group of particles except in the case of the protons from the reaction.

* $^{7}_{3}Li$: residual nucleus left in the excited level at 440 kev.

† $^{17}_{8}O$: residual nucleus left in the excited level at 850 kev.

Confirmation for the correctness of the above interpretation of the results is provided by similar observations in which we accept for measurement either only α-particles or only protons. The difference in the characteristics of the tracks of the two types of particles is well displayed in plate 1, where typical α-particles and protons are indicated. In figure 2(*b*) we show the results of observations in which we accept only α-particles for measurement. It will be seen that group (iii), which we have attributed to protons, has completely disappeared.

The results of independent observations in which protons only are accepted for measurement are represented in figure 2(*c*). The peak, (viii), in this figure occurs at a range in the middle of the continuous distribution of α-particles. It corresponds well with protons from the reaction $^{16}_{8}O + ^{2}_{1}H \rightarrow ^{17}_{8}O + ^{1}_{1}H$, the $^{17}_{8}O$ nuclei being formed in the excited state at 0·85 Mev. (Cockcroft and Lewis, 1936 b).

The results of the measurements of the mean range of the particles constituting the different groups, the reactions to which we attribute them, the corresponding Q values and the calculated mean energy of the particles, are tabulated in table 1. We use the values of Q obtained in the most accurate measurements, references being given to the sources from which they are taken. d is the probable error in the determination of the mean range of a group due to statistical fluctuations only. The calculated values of the energy take into account the thick-target corrections as treated by Bethe and Livingston (1937).

§ 5. EXPERIMENTS WITH BERYLLIUM

Because of the thinness of the beryllium film and its resistance to corrosion under the impact of the beam, the plate obtained with this target provided the most satisfactory conditions for measurement. A typical photo-micrograph of 0·1 mm^2 of the surface of the exposed plate is shown in plate 2. Owing to the relatively low exposure, the number of disintegration particles per unit area is small compared with that obtained with the lithium target, and the most prominent feature in the photograph is the large number of short tracks. These are due to the elastically scattered primary deuterons and the nuclear recoil particles formed in the various disintegration processes.

The range distribution obtained from the measurement of 2000 tracks is shown in figure 3, the reactions giving rise to the particles which produce the different peaks being tabulated with other details in table 2.

The α-particles from the bombardment of beryllium by deuterons have been studied in detail by Graves (1940). These authors found that the distribution in energy of the particles is complex, corresponding to the formation of the product nuclei, $^{7}_{3}Li$, either in the ground state or in an excited state of energy between 400 and 500 kev. The number of particles measured in our experiments is not sufficiently great to enable us to separate the two groups. The mean range of the observed distribution, peak (ii), figure 3, agrees well with the assumption that the most probable mode of disintegration is that giving rise to the $^{7}_{3}Li$ nuclei in the well known excited state at 440 kev.

The particles giving rise to peak (iii) produce tracks with a grain-spacing characteristic of protons and the mean range of the groups corresponds to an energy of 1·69 ± 0·8 Mev. Of the light elements which might be present as contaminants in the beryllium target, the only one which could give rise to a group of protons of about this energy is $^{16}_{8}O$. Protons of energy 1·81 Mev. could then be produced in the reaction $^{16}_{8}O + ^{2}_{1}H \rightarrow ^{17}_{8}O + ^{1}_{1}H$, the $^{17}_{8}O$ nuclei being left in the excited state at 0·85 Mev. If the observed proton group were produced by this reaction we should expect, however, to find a second group, of about the same intensity, corresponding to the formation of $^{17}_{8}O$ in the ground state. There is no trace of such a group, and we can conclude that there is no appreciable oxygen contamination of the target. We therefore assume that the observed protons are produced in the reaction $^{9}_{4}Be + ^{2}_{1}H \rightarrow ^{10}_{4}Be + ^{1}_{1}H$, the $^{10}_{4}Be$ nucleus being formed in an excited state. On this assumption, the energy released in the reaction is 1·11 Mev., a value which

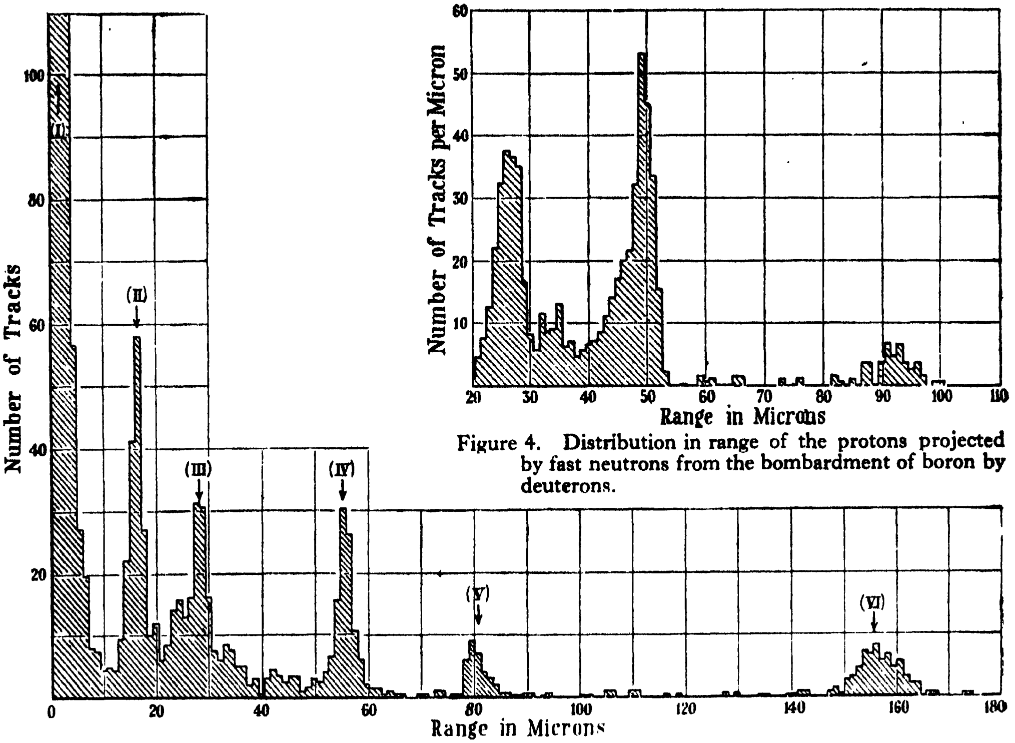


Figure 4. Distribution in range of the protons projected by fast neutrons from the bombardment of boron by deuterons.

Figure 3. Distribution in range of particles from the bombardment of beryllium by 900 kev. deuterons.

Table 2

Reaction	Q (Mev.)	E (Mev.)	R (microns)	d (microns)	Ref.
(i) Elastically scattered deuterons and recoil nuclei from (ii) and (iii)					
(ii) ${}^9_4\mathrm{Be}+{}^2_1\mathrm{H}\rightarrow{}^{*7}_3\mathrm{Li}+{}^4_2\mathrm{He}$	$6{\cdot}65\pm0{\cdot}02$	$4{\cdot}64\pm0{\cdot}02$	18·2	0·10	(1)
(iii) ${}_4\mathrm{Be}+{}^2_1\mathrm{H}\rightarrow{}^7_3\mathrm{Li}+{}^4_2\mathrm{He}$	$7{\cdot}09\pm0{\cdot}02$	$4{\cdot}92\pm0{\cdot}02$	—	—	(1)
(iv) ${}^9_4\mathrm{Be}+{}^2_1\mathrm{H}\rightarrow{}^{\dagger 10}_4\mathrm{Be}+{}^1_1\mathrm{H}$	$(1{\cdot}11\pm0{\cdot}08)$	$(1{\cdot}66\pm0{\cdot}08)$	29·8	0·1	
(v) ${}^9_4\mathrm{Be}+{}^2_1\mathrm{H}\rightarrow{}^8_4\mathrm{Be}+{}^3_1\mathrm{H}$	$4{\cdot}32\pm0{\cdot}05$	$3{\cdot}63\pm0{\cdot}04$	57·8	0·10	(2), (3)
(vi) ${}^2_1\mathrm{H}+{}^2_1\mathrm{H}\rightarrow{}^3_1\mathrm{H}+{}^1_1\mathrm{H}$ or ${}^{12}_6\mathrm{C}+{}^2_1\mathrm{H}\rightarrow{}^{13}_6\mathrm{C}+{}^1_1\mathrm{H}$	$3{\cdot}08\pm0{\cdot}02$ $2{\cdot}71\pm0{\cdot}05$	$3{\cdot}20\pm0{\cdot}02$ $3{\cdot}25\pm0{\cdot}02$	82·7	1·2	(2)
(vii) ${}^9_4\mathrm{Be}+{}^2_1\mathrm{H}\rightarrow{}^{10}_4\mathrm{Be}+{}^1_1\mathrm{H}$	$4{\cdot}51\pm0{\cdot}10$	$4{\cdot}75\pm0{\cdot}10$	159·5	0·2	(2), (3)

(1) Graves (1940).
(2) Oliphant, Kempton and Rutherford (1935 b).
(3) Williams, Haxby and Shepherd (1937 b).

For the reaction (vi) we adopted the mean value of Q given in references (2) and (3).

* ${}^7_3\mathrm{Li}$ left in the well known excited level of energy 440 kev.

† Values of Q and E, for reaction (iii), are calculated assuming that the proton group corresponds to a ${}^9_4\mathrm{Be}+{}^2_1\mathrm{H}\rightarrow{}^{10}_4\mathrm{Be}+{}^1_1\mathrm{H}$ reaction in which ${}^{10}_4\mathrm{Be}$ is left in an excited level.

corresponds to an excited state of energy 3·40 Mev. in $^{10}_{4}Be$. This value is deduced from the mass of Be in the ground state corresponding to the mean of the values for the energy released in the reaction $^{9}_{4}Be + ^{2}_{1}H \rightarrow ^{10}_{4}Be + ^{1}_{1}H$, determined by Oliphant, Kempton and Rutherford (1935 b) and by Williams, Shepherd and Haxby (1937 a), viz. 4·50 Mev. Alternatively, we can assume the particles to be tritons corresponding to the formation of $^{8}_{3}Be$ in an excited state. The corresponding value of the energy of this state is then found to be 2·2 Mev. Since no such state has been observed and since, in addition, we should expect the group to be inhomogeneous owing to the rapid disintegration of the $^{8}_{3}Be$ nuclei, we can reject this interpretation of the results.*

We have mentioned that most of the short-range particles are due to elastically scattered deuterons with a continuous distribution in range. In addition, however, we may expect to observe $^{7}_{3}Li$ recoil nuclei of energy 2·95 Mev. produced in the reaction $^{9}_{4}Be + ^{2}_{1}H \rightarrow ^{7}_{3}Li + ^{4}_{2}He$, and we believe that the discriminating power of the emulsion allows us to distinguish such tracks, in some cases, from the deuterons of approximately the same range. Characteristic tracks of tritons constituting group (iv) and protons from groups (v) and (vi) are indicated in plate 2. It is not possible, in our experience, to distinguish by inspection the tracks of tritons from those of α-particles of the same range in the B_1 emulsion.

§ 6. EXPERIMENTS WITH BORON

The exposures obtained with a boron target were not so satisfactory as those with beryllium and lithium because emulsions covered with a gelatine supercoat, 2 μ thick, were employed. Such a supercoat is designed to prevent surface marks from pressure or light friction, and the plates were produced for work on neutron spectra where the protons recoil start in the body of the emulsion. For experiments with fast-charged particles, the use of such plates leads to the difficulty that the beginnings of the tracks are not clearly defined, with a consequent loss of precision in the range measurements.

We may point out that the application of such a supercoat is also not entirely satisfactory for work with neutrons. Our preliminary observations indicate that there is a diffusion of silver halide grains into the supercoat so that the surface of the emulsion proper tends to be ill defined. As a result it becomes difficult to decide, in some cases, whether a track has left the emulsion, and such an ambiguity is serious in experiments where the "escape" of tracks has to be taken into account.

In spite of this difficulty we have made measurements of the lengths of the long protons produced in the reaction $^{10}_{5}B + ^{2}_{1}H \rightarrow ^{11}_{5}B + ^{1}_{1}H$, which are grouped in three peaks, corresponding to the formation of $^{11}_{5}B$ nuclei in the ground state or in excited states at 1·5 and 3·4 Mev. (Cockcroft and Lewis, 1936 a). The results of the measurements of 400 tracks are tabulated in table 3.

Plate 3 shows photo-micrographs of two interesting rare events observed in the course of the measurements of the plates obtained with the boron target. (*a*) is a proton-proton collision; (*b*), a photograph of the same event, was obtained by inclining the plate on the microscope stage so that the two branches could be observed in focus at the same time, and it established the fact that the three component tracks are co-planar. The poor optical conditions, due to the inclined

---

* In a letter to *Nature* (Lattes, Fowler and Cuer, 1947) the excited state of $^{10}_{4}Be$ was incorrectly stated to be 2·2 Mev. instead of 3·4.

plate, result in a photograph of inferior quality, but it may be noticed that, even with such a large tilt in the subject, the essential details of the collision can be well displayed.

The angle between the two arms of the tracks as measured in the photograph is 86°. This apparent degree of departure from a right angle, 4°, is greater than that which must have existed when the particles were produced in the emulsion, before development, but when the correction for shrinkage is applied a discrepancy of about 3° remains. This may be attributed to a small-angle scattering of one of the protons within one or two microns of the point of collision. Such small-angle Coulomb scattering is very frequent at low energies in the N.R. emulsions because of the high concentration of silver and bromine.

Table 3

Reaction	Q (Mev.)	E (Mev.)	R (microns)	d (microns)	Ref.
(i) ${}^{10}_{5}B + {}^{2}_{1}H \rightarrow {}^{*}{}^{11}_{5}B + {}^{1}_{1}H$	4·71	4·99	171	2·5	(1), (2)
(ii) ${}^{10}_{5}B + {}^{2}_{1}H \rightarrow {}^{\dagger}{}^{11}_{5}B + {}^{1}_{1}H$	7·00	7·09	320	3·0	(1), (2)
(iii) ${}^{10}_{5}B + {}^{2}_{1}H \rightarrow {}^{11}_{5}B + {}^{1}_{1}H$	$9·14 \pm 0·06$	$9·05 \pm 0·06$	474	3·0	(1), (2)

(1) Cockcroft and Lewis (1936).
(2) Livingston and Bethe (1937).
Ranges corrected for " thick target " effect.
*† : ${}^{11}_{5}$B, residual nuclei left in 2·14 and 4·13 Mev. excited levels respectively.

The second event, shown in (*c*), is a large-angle deflection of a proton by collision with a nucleus of silver or bromine. If we assume that it was produced by an elastic collision, the energy of the proton at the point of scattering was 4 Mev. Such large-angle deflections are rare with protons of this energy.

§ 7. EXPERIMENTS WITH NEUTRONS*

In order to determine the range of protons with energies greater than the values available in the previous reactions, measurements have been made on plates exposed to fast neutrons from the reaction ${}^{11}_{5}B + {}^{2}_{1}H \rightarrow {}^{12}_{6}C + {}^{1}_{0}n$, the exposure being made in the manner described by Powell (1943). In examining the plates, only protons projected at angles less than 8° with the direction of the incident neutrons were accepted for measurement.

From the approximate value of the range-energy relation we can determine the energy E' of the proton projected at an angle θ. The energy E_0 of the primary neutron producing the recoil track can then be found together with ΔR, the amount by which the observed range of the recoil track is less than the value to be expected for head-on collision, $\theta = 0$. The corrected values so obtained are used to obtain the distribution in range shown in figure 4, which displays the three well known groups for which the values of the mean neutron energy are approximately 6, 9 and 13 Mev. respectively in the conditions of our exposure. The values of the mean ranges of these groups and other particulars are shown in table 4.

---

* We are indebted to Mrs. Andrews for the measurements in § 7.

§8. THE RANGE-ENERGY RELATION FOR PROTONS

From the values of the energy and mean range of the different groups of particles given in tables 1 to 4 we can determine a range-energy curve for protons in the

Table 4

Reaction	*Q* (Mev.)	*E* (Mev.)	*R* (microns)	*d* (microns)	Ref.
(i) $^{10}B + ^{2}_{1}H \rightarrow ^{11}_{6}C + ^{1}_{0}n$	6·08	6·24	256	1·00	(1), (2)
(ii) $^{11}_{5}B + ^{2}_{1}H \rightarrow {}^{*}{}^{12}_{6}C + ^{1}_{0}n$	9·10	9·09	482	0·83	(1), (2)
(iii) $^{11}_{5}B + ^{2}_{1}H \rightarrow ^{12}_{6}C + ^{1}_{0}n$	13·40±0·30	13·06±0·30	896	3·6	(1), (2)

(1) Bonner and Brubaker (1936).
(2) Livingston and Bethe (1937).
Ranges corrected for " thick target " effect.
* $^{12}_{6}C$: residual nuclei left in 4·4 Mev. excited level.

Table 5

Reaction	*E* (Mev.)	*R* (cm. air)	*r* (*μ* in emulsion)	*d*	*R*/*r*
$^{11}_{5}B + ^{2}_{1}H \rightarrow ^{12}_{6}C + ^{1}_{0}n$	13·06±0·30	185·0	896·1	4·0	2064±9
$^{11}_{5}B + ^{2}_{1}H \rightarrow {}^{*}{}^{12}_{6}C + ^{1}_{0}n$	9·09	97·0	482·0	1·0	2012±5
$^{10}_{5}B + ^{2}_{1}H \rightarrow ^{12}_{5}B + ^{1}_{1}H$	9·05	96·2	474·0	3·0	2029±15
$^{10}_{5}B + ^{2}_{1}H \rightarrow {}^{*}{}^{11}_{6}C + ^{1}_{0}n$	6·24	50·0	256·0	1·0	1953±8
$^{10}_{5}B + ^{2}_{1}H \rightarrow ^{11}_{5}B + ^{1}_{1}H$	7·09	62·6	321·0	3·0	1950±20
$^{10}_{5}B + ^{2}_{1}H \rightarrow {}^{*}{}^{11}_{5}B + ^{1}_{1}H$	4·99	33·8	171·0	2·5	1977±28
$^{6}_{3}Li + ^{2}_{1}H \rightarrow ^{7}_{3}Li + ^{1}_{1}H$	4·96±0·12	33·5	170·0	0·2	1970±2
$^{9}_{4}Be + ^{2}_{1}H \rightarrow ^{10}_{4}Be + ^{1}_{1}H$	4·75±0·10	31·0	159·5	0·2	1943±3
$^{6}_{3}Li + ^{2}_{1}H \rightarrow {}^{*}{}^{7}_{3}Li + ^{1}_{1}H$	4·51±0·12	28·4	147·0	0·2	1931±2
$^{2}_{1}H + ^{2}_{1}H \rightarrow ^{3}_{1}H + ^{1}_{1}H$	3·20±0·02	15·7	82·7	0·2	1898±5
$^{12}_{6}C + ^{2}_{1}H \rightarrow ^{13}_{6}C + ^{1}_{1}H$	3·25±0·05	16·0	83·9	0·1	1907±3
$^{16}_{8}O + ^{2}_{1}H \rightarrow ^{17}_{8}O + ^{1}_{1}H$	2·59±0·06	11·0	59·6	0·3	1845±9
$^{9}_{4}Be + ^{2}_{1}H \rightarrow ^{8}_{3}Be + ^{3}_{1}H$	3·63±0·04	9·3	57·8	0·1	1606
Proton of same velocity	1·20±0·02	3·1	19·3	0·04	1606±15

R denotes range in air in cm. ; *r* denotes range in the emulsion in microns.

Assuming that peak (v), figure 2, obtained in the lithium exposure is due to carbon contamination of the target.

d refers only to the " probable error " in measuring the baricentre of a group and gives only the statistical weight of the point. Other sources of error may arise from the calibration and from uncertainties in the value of *Q*.

emulsion. The results are summarized in table 5 and shown diagrammatically in figure 5 (*a*) and 5 (*b*). From the range-energy curve in air given by Bethe we can determine the variation of the stopping power of the emulsion relative to air, and the results are represented in figure 6. It will be seen that there is a considerable variation of the mean stopping power as the range of the protons increases. The point at 1·2 Mev. is deduced from the observed mean range of the tritons produced in the reaction ${}^{9}_{4}\mathrm{Be}+{}^{2}_{1}\mathrm{H}\rightarrow{}^{8}_{3}\mathrm{Be}+{}^{3}_{1}\mathrm{H}$, for which the release of energy, Q, has been determined by Williams, Haxby and Shepherd (1937 b), $Q=4{\cdot}32\pm0{\cdot}05$ Mev. The calculated energy of the tritons emitted at 80° with the direction of the primary deuterons, of energy 900 kev., is $3{\cdot}63\pm0{\cdot}05$ Mev. If we assume that the range of a triton is three times that of a proton of the same velocity, we obtain the value given in table 5 for the range of protons of energy 1·21 Mev.

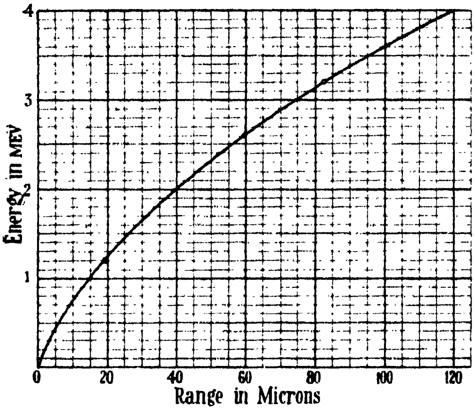


Figure 5 (*a*). Range-energy curve for protons. Nuclear research emulsion, type B1.

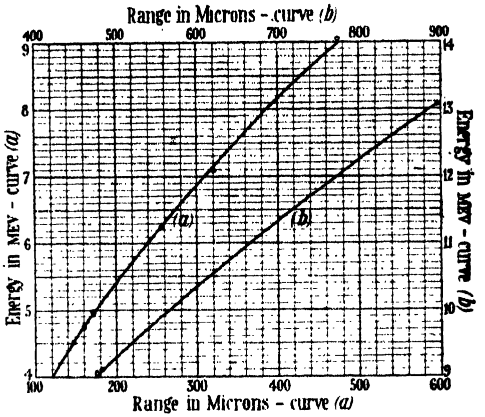


Figure 5 (*b*).

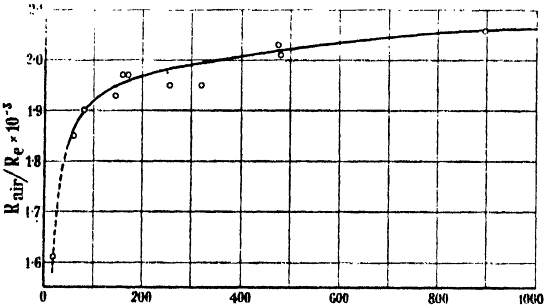


Figure 6. Mean stopping power of emulsion for protons of different ranges.

Although out present measurements are confined to the emulsion of type B1, we may assume that they will apply to the other types, after small corrections, since the manufacturers aim at producing emulsions of constant atomic composition. There may, however, be small changes in stopping power in plates in different batches, although we have not yet detected such an effect.

§ 9. RANGE-ENERGY RELATION FOR α-PARTICLES

The range-energy relation for α-particles has been determined, up to an energy of 9 Mev., by measurements of the lengths of the tracks produced by the homogeneous groups of particles emitted by the natural radioactive substances. The value at 13 Mev. has been deduced by employing the α-particles from the reaction ${}^{6}_{3}\mathrm{Li}+{}^{2}_{1}\mathrm{H}\rightarrow{}^{4}_{2}\mathrm{He}+{}^{4}_{2}\mathrm{He}$.

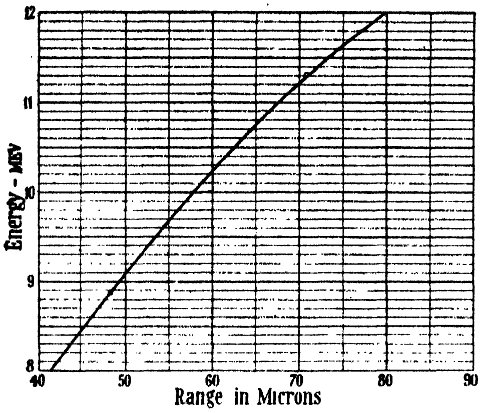


Figure 7 (*a*). Range-energy curve for α-particles. Nuclear research emulsion, type B1.

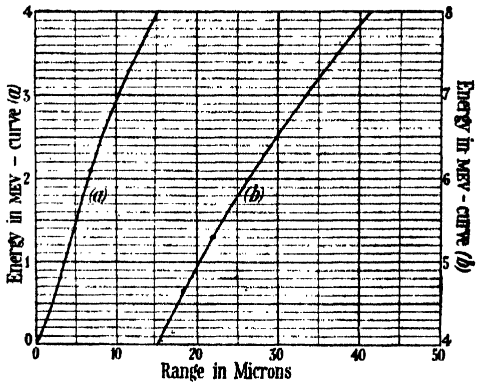


Figure 7 (*b*).

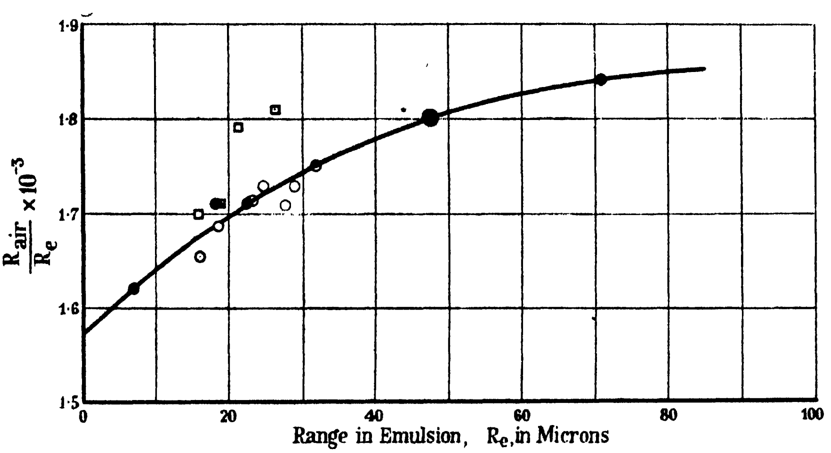


Figure 8. Stopping power for α-particles.

● Present experiments (B1).

☐ Tsien, Chastel, Faraggi and Vigneron, 1946 (A1).

○ D. L. Livesey and Green—private communication (C2).

Table 6

Reaction	E (Mev.)	Ranges in air	Ranges in emulsion	$\frac{R_a}{R_e} \times 10^{-3}$
		R (cm.)	R (2)	
${}^6_3\text{Li} + {}^2_1\text{H} \rightarrow {}^4_2\text{He} + {}^4_2\text{He}$	11·32±0·02	13·04	70·8 ±0·2	1·842±0·005
ThC′ α-particles	8·78	8·57	47·5 ±0·05	1·804±0·002
Po α-particles	5·30	3·84	22·5 ±0·06	1·707±0·003
${}^9_4\text{Be} + {}^2_1\text{H} \rightarrow {}^{*7}_{3}\text{Li} + {}^4_2\text{He}$	6·64	3·12	18·2 ±0·1	1·715±0·010
Sm α-particles	2·1	1·13	6·95±0·05	1·628±0·012

Table 6 gives a summary of the results used in drawing the curves shown in figures 7 (*a*) and 7 (*b*). We are indebted to L. L. Green, W. M. Gibson and D. L. Livesey for permission to include the results of their measurements on the α-particles from U I, U II, ThC and ThC′. Figure 8 gives the variation of the stopping power of the emulsion with the range of the particles.

§ 10. RESOLVING POWER

In order to display the resolving-power of the method, we show in figures 9 and 10 the apparent distribution in energy of the tracks produced by typical homogeneous distribution in energy tritons and α-particles. The curves are deduced from the observed distribution in range of the tracks corresponding to the different groups, using the appropriate range-energy curve, and a summary is given in table 7. The most significant results are those obtained with a beryllium target which was exceptionally thin. In this case the observed width of the proton group is only 25% greater than that expected from straggling. In the other cases, the observed width can be accounted for in terms of the thickness of the targets

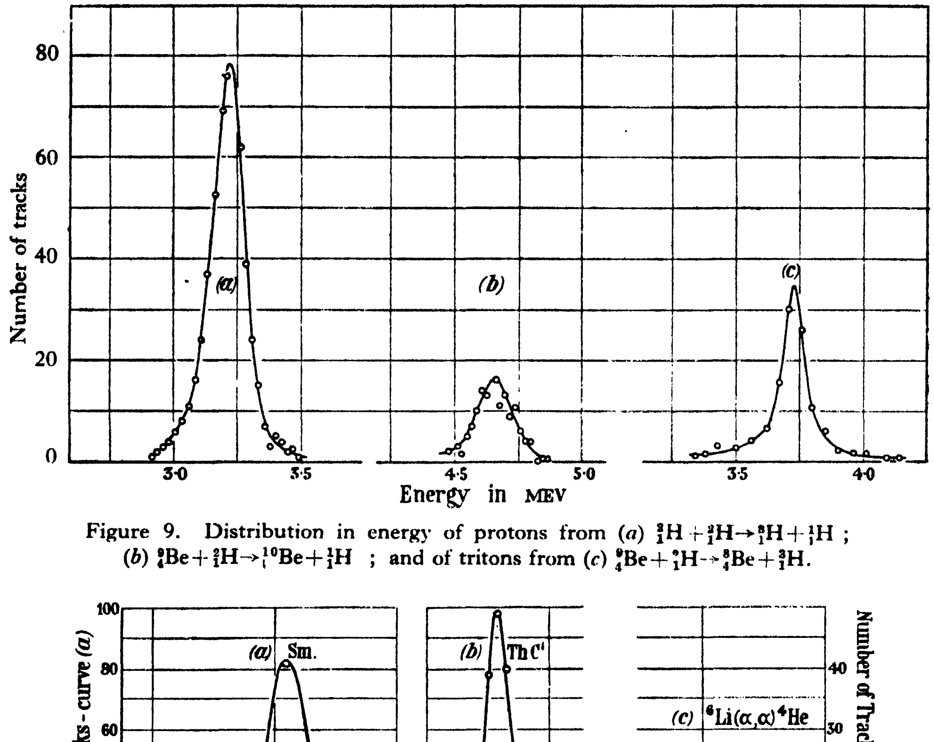


Figure 9. Distribution in energy of protons from (*a*) $^2_1H + ^2_1H \rightarrow ^3_1H + ^1_1H$; (*b*) $^9_4Be + ^2_1H \rightarrow ^{10}_4Be + ^1_1H$; and of tritons from (*c*) $^9_4Be + ^2_1H \rightarrow ^8_4Be + ^3_1H$.

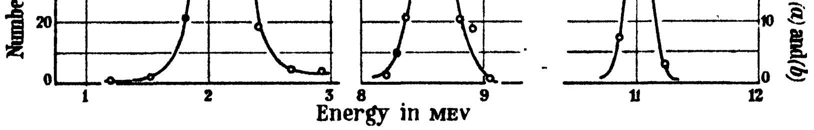


Figure 10.

employed. We may conclude that for protons in the interval from 2 to 13 Mev. the uncertainty in the energy of a particle, as deduced from its range in an emulsion, is not appreciably greater than that due to straggling.

Table 7

Reaction	E (Mev.)	$R_e(\mu)$	S (%)	S' (%)	Remarks
		Protons			
$^9_4Be + ^2_1H \to ^{10}_4Be + ^1H$	4·75	159·5	1·96	2·4	
$^2_1H + ^2_1H \to ^3_1H + ^1_1H$	3·20	82·7	2·05	2·5	Beryllium target
$^2_1H + ^2_1H \to ^3_1H + ^1_1H$ $^{12}_6C + ^2_1H \to ^{13}_6C + ^1_1H$	3·20	83·9	2·05	5·75	Probably due to $^{12}_6C + ^2_1H \to ^{13}_6C + ^1_1H$ Lithium target
$^9_4Be + ^2_1H \to ^8_4Be + ^3_1H$	3·63	57·8		4·8	
$^{11}_5B + ^2_1H \to ^{12}_6C + ^1_0n$	13·06	896·0	1·8	3·3	Thick target
$^{11}_5B + ^2_1H \to {}^{*12}_{6}C + ^1_0n$	0·99	482·0	1·7	3·3	Thick target
		α-particles			
$^6_3Li + ^2_1He \to ^4_2He + ^4_2He$	11·32	70·8	1·05	3·5	Thick target
ThC′ α-particles	8·78	47·5	1·07	2·7	
Po α-particles	5·30	22·0	1·17	4·0	
Sm α-particles	2·1	6·95		14·0	

S = theoretical straggling in air (Bethe and Livingston, 1937).

S' = " experimental straggling " = $\sqrt{\frac{\pi}{2}} \times \sqrt{\Sigma_n \frac{(x-x_i)^2}{n}}$, where x_i is the range corresponding to the " baricentre " of the group, x the range of a particular track and n the number of tracks composing the " group ".

The corresponding results for α-particles are shown in figure 10. They indicate that, because of the much less favourable range-energy relation, the uncertainty in the true range, due to the finite size of the silver halide grains, leads to errors in the estimated energy of the particles, which are much greater than those due to straggling. This effect is especially marked at low energies, so that the peak of greatest "width" is that obtained in experiments with the α-particles from samarium. As the energy of the α-particles increases, the "half-width" becomes smaller. We must anticipate that with particles of progressively greater energies than those of which we have experience, the half-width will reach a minimum and then increase again as the errors due to the finite grain size become small compared with those due to straggling. For work at moderate energies it is therefore desirable to reduce the errors due to "apparent straggling" by working with emulsions of smaller grain size such as are provided by the C, E and D series of emulsions. Such an improvement in precision would seem to be particularly important in measurements of the multiple disintegration phenomena which we meet in the disintegration of nuclei produced by the high-energy γ-rays from the betatron or the cosmic rays (Heitler, Powell and Fertel, 1939; Heitler, Powell and Heitler, 1940; Baldwin and Klauber, 1946).

ACKNOWLEDGMENTS

The present work was carried out in the H. H. Wills Physical Laboratory, University of Bristol, and we wish to express our appreciation to the Director of the laboratory, Prof. A. M. Tyndall, for much kindness and hospitality. The experiments were made at the suggestion of Dr. C. F. Powell and under his direction. We had the advantage of being able to make full use of the techniques developed by him and Dr. G. P. S. Occhialini, and of the special facilities of the laboratory for the application of the photographic method to nuclear physics. The authors wish to express their gratitude to them for their continuous interest and active help.

We are indebted to Sir Lawrence Bragg, Dr. Burcham and Mr. Livesey of the Cavendish Laboratory, who kindly made it possible for us to obtain the exposures on the Cambridge High Tension set, and to several of our colleagues in the laboratory for assistance on many occasions, especially to Mr. R. M. Payne and Mr. E. Samuel. We are also indebted to W. M. Gibson, L. L. Green and D. L. Livesey for permission to include in table 6 the results of their measurements on the α-particles from Ui and Uii and from the Th radioactive series.

One of us (C. M. G. L.) was enabled to undertake the work by a grant from the University of Bristol and from the Getulio Vargas Foundation.

REFERENCES

Baldwin and Klaiber, 1946. *Phys. Rev.*, **70**, 259.
Bonner and Brubaker, 1936. *Phys. Rev.*, **50**, 781.
Cockcroft and Lewis, 1936 a. *Proc. Roy. Soc.*, A, **154**, 246 ; 1936 b. *Ibid.*, **154**, 261.
Graves, 1940. *Phys. Rev.*, **57**, 855.
Guggenheimer, Heitler and Powell. In the press.
Heitler, Powell and Fertel, 1939. *Nature, Lond.*, **144**, 283.
Heitler, Powell and Heitler, 1940. *Nature, Lond.*, **146**, 65.
Lattes, Fowler and Cuer, 1947. *Nature, Lond.*, **159**, 301.
Livingston and Bethe, 1937. *Rev. Mod. Phys.*, **9**, 326.
Oliphant, Kempton and Rutherford, 1935 a. *Proc. Roy. Soc.*, A, **149**, 406; 1935 b. *Ibid.*, **150**, 241.
Powell, 1943. *Proc. Roy. Soc.*, A, **181**, 344.
Powell, Occhialini, Livesey and Chilton, 1946. *J. Sci. Instrum.*, **23**, 102.
Rumbaugh, Roberts and Hafstad, 1938. *Phys. Rev.*, **54**, 657.
Smith, 1939. *Phys. Rev.*, **56**, 548.
Williams, Haxby and Shepherd, 1937 a. *Phys. Rev.*, **52**, 390 ; 1937 b. *Ibid.*, **52**, 1031.

DISCUSSION

on paper by E. F. DALY and G. B. B. M. SUTHERLAND entitled "An Infra-red Spectroscope with Cathode-ray Presentation" (*Proc. Phys. Soc.*, **59**, 77 (1947)).

Dr. F. AUGHTIE. Plate 1 of this paper shows two displays: (*a*) without and (*b*) with smoothing by a resistance capacity filter stated to be of about 2 cycles/sec. width. No reference is made in the paper to the time delay of the filter; this will be of the order of the reciprocal of the band width and will cause the appropriate vertical deflection of the trace to occur at a position displaced horizontally by a distance corresponding to this time interval. It is quite noticeable that several distinctive features of the trace of plate 1 (*b*) are displaced to the left by 2–3 mm. from the corresponding features of 1 (*a*). This displacement is of the order to be expected from the band and sweep frequency employed.

Thus if the smoothed curves are to be employed, the wave length scale should be displaced by a distance corresponding to the time delay. Preferably the cam should be arranged to operate an auxiliary shutter which would momentarily interrupt the light at known settings, so that the wave length scale appropriate to the sweep-speed in use can be determined from a blank observation.

The effect of filter delay would, of course, be more serious if a faster rate of scan were used.

AUTHORS' reply. We are grateful to Dr. Aughtie for drawing attention to this point, which, perhaps, ought to have been more fully explained in our paper. We are, of course, well aware of the distortion both in amplitude and in phase introduced by the narrow band-resistance capacity filter, and we did remark that its use is accompanied by a sacrifice in the precision with which spectra are displayed. We do not feel, however, that this small displacement in the trace in going from one type of presentation to the other is important because in practice calibration is always done by reference to the spectra of compounds having sharp absorption bands of which the wave length has been determined accurately by grating measurements. Consideration of the filter delay is thus unnecessary provided the calibration spectra are displayed in the same way as the spectra under consideration.

CORRIGENDA

"The lines of force through neutral points in a magnetic field", by DAVID OWEN (*Proc. Phys. Soc.*, **59**, 14 (1947)).

Page 17. The words " this case is illustrated in figure 3 " should be deleted.

" Theory of the proton synchrotron ", by J. S. GOODEN, H. H. JENSEN and J. L. SYMONDS (*Proc. Phys. Soc.*, **59**, 677 (1947)).

Page 680, eleventh line from foot of page. The equation should read

$$(\phi_2+\phi_1-\pi)\sin\phi_1+\cos\phi_2+\cos\phi_1=0.$$

Page 686, equation (5). The numerical constant (3·3) on the right-hand side should be replaced by (0·53).

Page 686, equation (7). R_0 should be R_0^2.

Page 686, equation (8). In the right-hand side the factor 2 in the numerator should be 6.

Page 690, equation (11). In the right-hand side the factor 2 in the numerator should be 8.

Page 691, equation (17) should read

$$B=\frac{ps\cot\phi_1\tan\theta}{2(s+1)(1-n)}.$$

REVIEWS OF BOOKS

The Physical Principles of Wave Guide Transmission and Antenna Systems, by W. H. WATSON. Pp. xiii + 208 with 95 figures. (Oxford: Clarendon Press, 1947.) 20*s*. net.

It may be said immediately that this is a good book. It does not always follow that one who is an expert original worker in a given field is also good at giving an account of the work in that field, but Professor Watson has succeeded in this case.

It is stated: " The aim of this book, which is addressed to physicists and engineers with theoretical interests, is to describe the way in which the technique of handling radio frequency transmission-lines has been extended to deal with propagation through hollow metal pipes known as wave guides." It is shown that simple circuit ideas, such as are applicable to ordinary transmission-line theory, may, to a large extent, be developed to treat many features of the propagation of the " dominant " wave in wave guides, as well as the loading of such guides by coupling and matching devices. Here the " dominant " wave is the one and only characteristic mode of propagation possible in a given wave guide when the dimensions of the latter are suitably chosen in relation to the frequency of the radiation concerned. Whilst in the main this is a theoretical treatise, the author has in mind the emphasis of the underlying physical principles, and he has succeeded admirably in his aim, namely, to bring out these principles in a clear and understandable manner.

Chapter I contains a treatment of the elements of wave propagation using the impedance concept. With the aid of the simple strip transmission-line it is shown how to transfer the circuit ideas of impedance and admittance to waves in space. As a result it is possible to discard the elaborate differential equations employed even with such simple waves, and to use instead algebra of no great complexity. The algebraic equations representing waves either on a transmission-line, or in space, can be easily connected with a geometrical representation in terms of the circle diagram. From this algebraic representation it follows that the propagation and loading of waves may be regarded as transformations which are found to be simple in type, and may be conveniently represented by matrices. This matrix algebra is developed in some detail, and the algebraic transformations necessary for dealing with various characteristics of plane wave propagation are considered.

Chapter II is devoted to a consideration of the dominant wave (H_{10} or TE_{10}) in a rectangular wave guide. This problem is attacked from the physical point of view, and the solution is obtained from the investigation of a system of two interfering trains of plane waves rather than by the more common procedure of simply applying Maxwell's equations and the boundary conditions—a procedure which does not lend itself so readily to a formation of a clear picture of the processes involved. Various aspects of the propagation of the dominant wave are dealt with: the energy flux, the current distribution on the walls of the wave guide, impedance, and the attenuation of the wave due to the finite conductivity of the walls.

A description is given in Chapter III of various measurement techniques which are necessary for an experimental investigation of the phenomena in wave-guide systems. This includes wave-meters, standing-wave detectors and power measurement: a short section deals with wave-guide stubs for impedance matching purposes. The treatment in this chapter is rather cursory and does no more than describe the underlying principles of the various instruments and techniques; it would not meet the needs of one who wished to design such equipment, but it should be stated that the author quite clearly does not aim at satisfying this requirement.

In Chapter IV there is a brief discussion of the general problem of multiple mode propagation of both E- and H-waves in rectangular and circular wave guides. The question of multiple propagation on the outside of a guide, which arises when an aperture is cut in the wall of the guide, is also mentioned, but is not developed to any great extent. The author then returns in Chapter V to a more detailed consideration of phenomena related to the propagation of the dominant wave in a rectangular wave guide, particularly in so far as the problem is affected by irises across the guide, and by coupling to the guide by slots cut in the walls. Babinet's principle is here extended to be of use in the microwave field. A restatement of the principle is necessary to deal with vector waves and reflecting screens. The iris is then considered as a grating problem. Bends in wave guides are discussed, and also the coupling of a wire antenna to the H_{10} wave in a rectangular guide: the impedance of such an antenna under these conditions is determined.

Chapters VI, VII and VIII cover the behaviour of slots cut in both the broad and narrow faces of a rectangular wave guide. Such characteristics as the coupling of these slots to the guide and their susceptance and conductance are determined. The various ways in which one guide may be coupled to another by suitably disposed slots in each are elaborated, and in general in these three chapters the theory is developed in some detail. A section is devoted to the consideration of antenna arrays using wave guides : this includes both arrays of linear dipoles energized by probes coupling through the broad face of the guide, and also arrays of slot radiators. The design of arrays to give various types of radiation pattern is discussed : this involves the determination of the distribution of conductances, amplitudes of excitation, and spacing of slots. Although in the main only linear arrays are considered, it is indicated also how two-dimensional wave-guide arrays may be constructed.

These three chapters are very well done : this is perhaps not surprising in view of the important contributions to the subject made by Professor Watson and his colleagues at McGill University during the late War.

In Chapter IX the author describes some further microwave devices, including the design of an array to produce a cosecant pattern, a Yagi array of slots, wave-guide switching systems and phase changers. Resonant cavities and the " Magic Tee " junction are dealt with in brief.

In the final chapter (X) Professor Watson considers field representations in connection with radiation and reception by antenna systems in wave guides. The basis of the analysis is the determination of expressions for the Hertz vectors of the field in a wave guide due (*a*) to an electric dipole, and hence to a current element, and (*b*) to a magnetic radiator which is realized by a distribution of tangential electric force on a surface where an aperture replaces part of a wave-guide wall. The thesis is developed in considerable detail, and such problems as those presented by an axis in a rectangular wave guide, radiation by an electric dipole and the impedance of such a linear antenna in a guide, and the characteristics of a resonant slot are analysed.

In general it may be said that the treatment is concise yet lucid throughout, and that Professor Watson has succeeded in bringing out clearly the physical principles involved in the transmission of waves through wave guides. He is to be congratulated on having filled so well a definite gap in the literature of this still quite new field. There is a minor criticism in relation to the title : at first glance one might imagine that the book was intended to deal with antenna systems in general, whereas, in fact, only some of those systems in which the wave guide plays an important part are discussed.

The book is well illustrated and produced, but it is a pity that so many references must still be to unpublished work. It is to be hoped that, in so far as it becomes possible, the author will endeavour to remedy this defect in any future editions. J. A. SAXTON.

The Metre-Kilogram-Second System of Electrical Units, by R. K. SAS and F. B. PIDDUCK. Pp. v+60. (London: Methuen and Co. Ltd., 1947). 4*s*. net.

A booklet to bring the metre-kilogram-second system of units to the notice of the scientific and technical public has been wanted in this country for several years, especially as the admirable publication of G. A. Campbell (*Bulletin* of the National Research Council, Washington, 1933, No. 93, p. 48) is not available here to a large circle of readers. For that reason the present publication is welcome and is bound to serve a useful purpose, if only to stimulate interest and discussion.

Put briefly, it happens that the orders of magnitude of the ampere and volt were so chosen that when the metre and kilogram are used as units of length and mass, the unit of power is the watt. This purely accidental coincidence is the main argument in favour of the MKS system. The authors are a trifle apologetic about the fact that in this system the density of water is 1000 kg. per m^3, but there should be no difficulty whatever there if the distinction between density and specific gravity is maintained. In the MKS system, as in other systems, the magnetic units can be chosen in two ways, which give rise to the normal and the so-called " rationalized " system respectively. The book under review, rightly we think, considers the latter system only. The main argument for a rationalized system is that it confines the factor 4π to formulae dealing with problems whose geometry is spherical, e.g., the potential due to a point charge, whereas the factor does not appear in problems of rectangular geometry, e.g., the capacitance of a parallel plate condenser.

The booklet is commendably short and can be read in an hour, but adequate care does not appear to have been given to its preparation, and the text does not seem to have been discussed with physicists intimately conversant with the subject. As a result a number of blemishes occur which may well produce the opposite of what the authors set out to achieve—the reader may be discouraged by finding too much complication in an inherently very simple system.

The proposal to use the word " pulse " for ω has no connection with the MKS system. In the past much thought has been given to this matter, but no entirely satisfactory solution appears to have been found. The proposal to call " aperture " what is generally called " solid angle " is also irrevelant.

These proposals are harmless, but that of the name " oersted " for the ampere-turn per metre is definitely in a different category: apart from the confusion its adoption would cause, it violates the excellent principle that special names should not henceforth be assigned to derived units when expression in terms of primary units is at the same time simple and indicative of the derivation. This rule applies to the ampere-turn per metre, whereas in the case of the volt-second per turn the first condition hardly applies, and accordingly the name "weber" has been proposed. The reviewer is not now connected with the national and international bodies dealing with nomenclature, but he ventures to express the hope that those responsible will resist any attempts to multiply christenings in the MKS family.

It is not clear how confusion would be avoided between permeability, i.e., ratio B/H, and relative permeability if the nomenclature proposed in the book were adopted. In the MKS system the symbol μ must be reserved for permeability, i.e., ratio B/H in any medium; μ_0 is then used to denote the corresponding quantity for a vacuum, and the ratio μ/μ_0, denoted, say, by μ_r, must be given a name other than just " permeability": the name *relative permeability* possesses the advantage of indicating the meaning of the symbol without further definition. The parallel case of permittivity should be similarly treated. This point is stressed here because lack of a well-thought-out nomenclature can cause pitfalls, instances of which are found on p. 45, line 2, where the ratio B/H for the conductor, μ, not μ_0, should have been used, and on p. 46, where B, and not H, should be equated to curl $\mathbf{A}$, in which $\mathbf{A}$ is the vector potential expressed in webers per metre.

The first sentence of p. 52, which contains the amazing statement that " Neumann's formula has a factor 10^{-7} " can only add to the " tribulations of the student " described in pp. 16–17, for the poor student had previously been induced to think that he would never again meet powers of 10 in his algebraic formulae. Neumann's formula contains nothing of the sort, and should be written

$$M=\frac{\mu}{4\pi}\int\frac{dl.dl'}{r},$$

where μ for a vacuum is 1·257 microhenry per metre.

The root of the trouble is that the authors, after giving on p. 23 the value of μ as $4\pi/10^7$, preserve this value explicitly in their formulae. Let them by all means explain that the ohm is chosen so that μ_0 has that value, but ever afterwards write this value as 1·257 microhenry per metre, in the same way as ϵ_0 on p. 41 is given as 8·854 micromicrofarads per metre. These are the values that " the reader must learn ", and not 10^{-7}. Incidentally, the opening sentences of pp. 38 and 51 are examples of the lack of care devoted to the preparation of the book. It may not be out of place here to recall to those wishing to employ or teach the MKS system the advisability of stating the units after formulae, e.g., $h=6{\cdot}62\times10^{-34}$ joule-seconds, $J=4185$ joules per kg.-calorie.

P. VIGOUREUX.

THE PROCEEDINGS OF
THE PHYSICAL SOCIETY

VOL. 59, PART 6 1 November 1947 No. 336

THE OPTICAL PROPERTIES OF AXIALLY SYMMETRIC MAGNETIC PRISMS

PART 1: THE STUDY OF RAYS IN A PLANE OF SYMMETRY, AND ITS APPLICATION TO THE DESIGN OF PRISM β-SPECTROSCOPES

By R. E. SIDAY,
Birkbeck College

MS. received 30 *January* 1947

ABSTRACT. The optical properties of the magnetic field in the neighbourhood of two equal co-axial magnetic poles of opposite polarity used as a magnetic prism are discussed. In this first part attention is confined to rays in the plane of symmetry between the poles over which the radial component of the field vanishes everywhere.

The focusing properties and aberrations are discussed generally, and detailed calculations are given :

(1) by simple analytical methods for the case of zero pole separation (homogeneous field);
(2) by numerical computation for the case of a pole separation equal to the radius of either pole.

Inferences are drawn concerning the use of such fields in β-spectroscopy, and it is shown that under certain conditions image formation by the prisms is optically corrected over a very wide aperture, or under other conditions can be corrected in compound instruments consisting of the prism and one or more magnetic lenses. Thus very high resolution is possible in such instruments. The overall optimum design for a β-spectroscope, however, cannot be made without a consideration of the focusing towards the plane of symmetry of rays confined to that plane. This focusing will be considered in Part 2.

§ 1. INTRODUCTION

THE magnetic prism has increasing applications in nuclear physics. Already in different forms of the mass spectrometer the magnetic prism plays a fundamental part, and in β-spectroscopy suggestions have been made regarding the use of prisms to obtain high resolving powers. Very little information, however, is available regarding the optical properties, particularly the aberrations, of these prisms.

This note will be confined to a consideration of the use as prisms of axially symmetric fields that possess a plane of symmetry in which the radial component of the field vanishes everywhere. In this first part only trajectories in the plane of symmetry will be considered, and the orbits discussed will be of the non-periodic type according to the classification of Coggershall and Muskat (1944), since for the cases considered the strength of the field will decrease as the radius increases.

It might be useful to describe more clearly what is meant by the term *prism* in this connexion. As will be discussed in detail, image formation by rays in the plane of symmetry occurs with varying quality according to the distance of the object from the axis of the field, and of course the usual method of semicircular focusing and its more recent modifications employ special cases of this image formation in which the object is immersed in the field. This paper will be concerned more with image formation when the object is so far from the axis that the field is negligibly strong, although a slight extension will be made in § 6 to the case in which the object is located where the field is weak but not negligible. There is no sharp distinction between semicircular focusing and the cases discussed here, although for the latter the general picture will be of rays entering the prism from a considerable distance from the axis, being deviated and finally emerging from a *virtual* focus. The focus will be virtual because the real image, in general, will occur where the field is not negligible and the rays will be further deviated in their passage away from the axis towards field-free space where they will belong to a virtual image.

In optics, the limit to resolution is imposed entirely by the finite wavelength of light. In the spectroscopy of charged particles, the limit arises entirely from the aberrations of image formation. It is clear that if the prism is to play a fundamental rôle in producing high resolution in any spectrometer, the aberrations of its image formation become of crucial importance.

Most mass spectrometers employ a sector-shaped prism of uniform field strength, the focusing properties of which were established by Herzog (1934). Aston, however, virtually uses a homogeneous axially symmetric field, and the optical properties of this for the particular case of this spectrometer were considered by Sawyer (1936). Coggershall and Muskat have examined ion paths in a wide variety of fields and have proposed new designs of mass spectrograph. In β-spectroscopy, however, the focusing properties of beams of much wider angle are important. Their properties will here be considered generally, and detailed calculations will be given for the two special cases (1) of a homogeneous field, (2) of the field in the neighbourhood of two perfect equal and opposite pole coaxial circular magnetic poles separated by a distance equal to the radius of either.

§ 2. GENERAL FOCUSING PROPERTIES IN THE PLANE OF SYMMETRY

Consider an electron moving in the plane of symmetry. Where the field is negligible at great distance from the axis, the trajectory is a straight line. Thus an electron entering the field of the prism moves initially in a straight line, is curved in the magnetic field, and finally emerges again along another straight trajectory. Because of the axial symmetry, the initial straight trajectory can be specified by a single parameter, *b*, which is the shortest distance of this straight line produced from the axis.

The differential equations of the motion of the electron of specific charge $\mathbf{e}/m$ moving in an axially symmetric field are

$$\frac{d^2r}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2 = -\frac{\mathbf{e}}{m} rH_z \frac{d\theta}{dt}, \qquad \ldots\ldots(1)$$

$$\frac{d}{dt}\left(r^2\frac{d\theta}{dt}\right) = -\frac{\mathbf{e}}{m} rH_r \frac{dz}{dt} + \frac{\mathbf{e}}{m} rH_z \frac{dr}{dt}, \qquad \ldots\ldots(2)$$

$$\frac{d^2z}{dt^2} = \frac{e}{m} rH_r \frac{d\theta}{dt}, \qquad \ldots\ldots(3)$$

where r, θ and z, are cylindrical coordinates defining the position of the particle with respect to the axis of the field and H_r and H_z are respectively the radial and axial components of the field at that point. The equations for a ray in the plane of symmetry where $H_r = 0$, reduce to

$$\frac{d^2r}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2 = -\frac{e}{m} rH_z \frac{d\theta}{dt}, \qquad \ldots\ldots(4)$$

$$\frac{d}{dt}\left(r^2 \frac{d\theta}{dt}\right) = \frac{e}{m} rH_z \frac{dr}{dt}. \qquad \ldots\ldots(5)$$

Integration of the last equation gives

$$\left| r^2 \frac{d\theta}{dt} \right|_1^2 = \frac{e}{m} \int_1^2 rH_z\, dr. \qquad \ldots\ldots(6)$$

If ψ is the angle between the trajectory at any point and the radius vector at that point, and if v is the constant velocity of the electron in the magnetic field

$$r^2 \frac{d\theta}{dt} = rv \sin\psi,$$

so that

$$r_1 \sin\psi_1 - r_2 \sin\psi_2 = \frac{1}{(H\rho)} \int_2^1 rH_z\, dr \qquad \ldots\ldots(7)$$

where $(H\rho)$ is the " momentum " mv/e. For very large values of r,

$$r_1 . \sin\psi = b.$$

Thus equation (7) becomes

$$r \sin\psi - b = \frac{1}{(H\rho)} \int_\infty^r rH_z\, dr, \qquad \ldots\ldots(8)$$

and this equation will subsequently be made the basis of ray tracing. During its passage through the field, the electron will follow a curved path which will at one point have a closest distance from the axis $r_{\min}$ whose value is obtained by putting $\psi = \pi/2$ so that

$$r_{\min} - b = \frac{1}{(H\rho)} \int_\infty^{r_{\min}} rH_z dr.$$

As H_z is a function of r only, the complete trajectory is symmetrical about this point of closest approach to the axis. In particular, the parameter b for the straight line along which the electron moves after it has left the field, has the same value as that of the straight trajectory before it entered the field of the prism. This can be formally shown, for if b and b' are the two values of these parameters respectively

$$b' - r_{\min} = -\frac{1}{(H\rho)} \int_\infty^{r_{\min}} rH_z dr,$$

$$r_{\min} - b = \frac{1}{(H\rho)} \int_\infty^{r_{\min}} rH_z dr,$$

so that $b = b'$.

Thus in a given field any trajectory is defined completely by the two parameters b and $(H\rho)$. The approaching electron of parameter b will be deviated through an angle $2.\phi_b$, which is equal to $\pi - 2\theta_b$ where θ_b is the total increase in the coordinate θ, as the electron approaches from infinity to the point of closest approach. These angles are marked in figure 1. In general, for a given $(H\rho)$, ϕ_b will depend on b, for which the following sign convention is adopted. If the length of the trajectory in the field is shorter than that of the trajectory directed initially towards the axis (i.e. with $b=0$), then b is taken as positive. If it is longer, b is negative. If $2.\phi_0$ is the deviation of a principal ray, that is, one initially directed towards the axis, and $2\phi_b$ that of a parallel ray of parameter b, after deviation, these two rays will appear to have come from a focus at a distance equal to

$$f_b = \frac{-b}{\sin(2\phi_b - 2\phi_0)} \qquad \ldots\ldots(9)$$

from the axis. If as an optical convention distances are measured from the origin O, and object distances are positive in a direction opposed to that of the motion of the electron, while image distances are positive in the direction of motion, then if f_b is to be positive for a convergent lens action, the sign of (9) is correct. If now f_b is constant for all values of b, the focus is free from aberration, but in any case the values of ϕ_b as a function of b give all the information required to examine the quality of the actual focus.

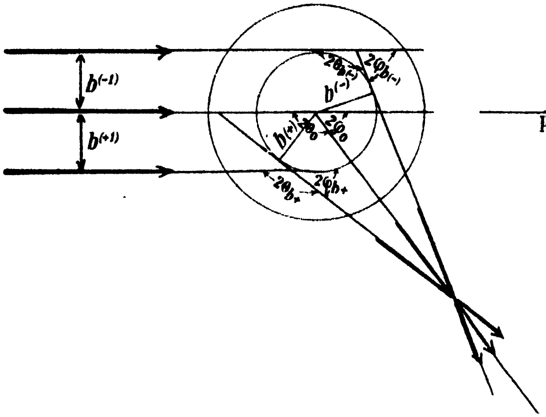


Figure 1. Focusing a parallel bundle.

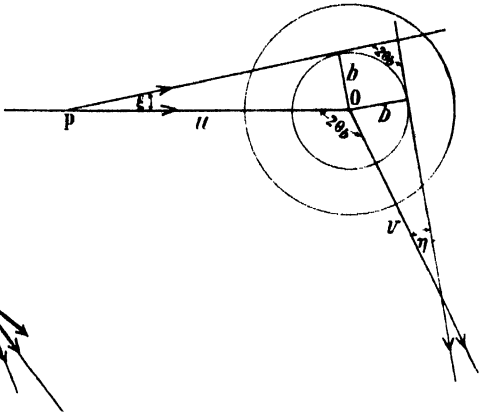


Figure 2. Imaging a point source.

In order to obtain the rules for the focusing of a point not at infinity, consider the two rays of figure 2 coming from the point P. Then $\sin\xi = b/u$, and $\sin\eta = b/v$, where u and v are object and image distances from the axis respectively.

$$\eta + 2\theta_0 + \xi = 2\theta_b,$$

$$\eta + \xi = 2\theta_b - 2\theta_0 = 2\phi_0 - 2\phi_b,$$

$$-\cos\eta\sin\xi - \sin\eta\cos\xi = \sin(2\phi_b - 2\phi_0),$$

$$\frac{\cos\eta}{u} + \frac{\cos\xi}{v} = \frac{1}{f_b}. \qquad \ldots\ldots(10)$$

The aperture defect shows up now in the presence of the cosine terms as well as in the general dependence of f_b on b. If as an approximation $\cos\eta$ is written as $1 - \frac{1}{2}b^2/v^2$, and $\cos\xi$ as $1 - \frac{1}{2}b^2/u^2$, then, to this order of accuracy, equation (10) becomes

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f_b\left(1 - \frac{1}{2}\frac{b^2}{uv}\right)}. \qquad \ldots\ldots(11)$$

If $f_b . (1 - \frac{1}{2}b^2/uv)$ is constant and equal to f_0 say, the image is free from aberration of this order. In general, both the terms f_b and $(1 - \frac{1}{2}b^2/uv)$ depend on b, and the possibility arises that for a certain value of u the two terms could cooperate to give corrected or partly corrected image formation. The variation of f_b with b will turn out in general to be too large to be essentially affected by the $(1 - \frac{1}{2}b^2/uv)$ term, but it will be shown that this is not always so, and that correction can in fact occur. For a given value of u, the conjugate value of v depends according to equation (11) on b, but in the correction term it will be sufficient to take for uv that derived from the approximate equation $\frac{1}{u} + \frac{1}{v} = \frac{1}{f_0}$. Then, as in elementary optics, the product uv is given by $f_0^2(1+m)(1+1/m)$ where m is the magnification, and the maximum effect of the term $(1 - \frac{1}{2}b^2uv)$ arises when uv is smallest, that is, when $m = 1$ and $u = v$.

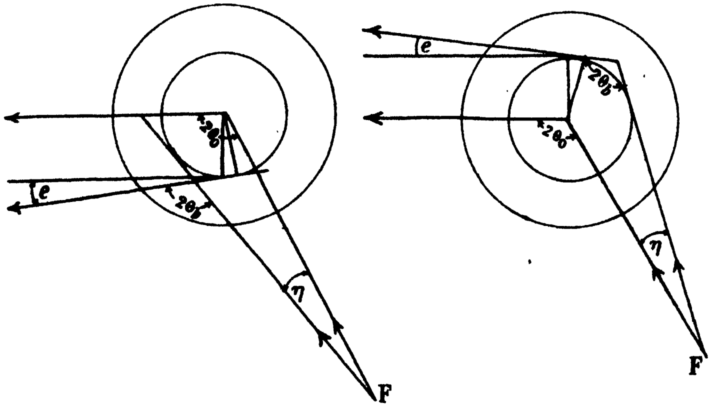


Figure 3. Sign of the aberration.

The condition for the absence of aberration to this order of accuracy can be expressed more conveniently as follows. If u is kept constant, and v_0 and v are the image distances for rays of parameters 0 and b respectively, then from (11)

$$\frac{1}{u} + \frac{1}{v_0} = \frac{1}{f_0},$$

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f_b}\left(1 + \frac{1}{2}\frac{b^2}{uv}\right),$$

$$\Delta\left(\frac{1}{v}\right) = \frac{1}{v_0} - \frac{1}{v} = \frac{1}{f_0} - \frac{1}{f_b} - \frac{1}{2}\frac{b^2}{uvf_0}.$$

$$\therefore -\Delta v = \frac{v^2}{f_0 f_b}\left(\overline{f_b - f_0} - \frac{1}{2}\frac{b^2 f_0}{uv}\right). \qquad \ldots\ldots(12)$$

Therefore for corrected image formation it is necessary that

$$f_b - f_0 = \tfrac{1}{2}\frac{b^2 f_0}{uv}, \qquad \ldots\ldots(13)$$

an equation that implies a quadratic form as the functional variation of f_b with b, as well as the correct magnitude of uv. Thus, in general, only partial correction can be expected, although in certain cases full correction does occur.

It will be useful also to consider sometimes the angular aberration. Figure 3 has been drawn in order to fix the magnitude and sign of the angular aberration in a parallel bundle. If this angle is e, then according to the sign of b, e is given by

$$\left.\begin{aligned} e &= \eta - (2\theta_b - 2\theta_0) \quad (b \text{ positive}),\\ e &= \eta - (2\theta_0 - 2\theta_b) \quad (b \text{ negative}), \end{aligned}\right\} \qquad \ldots\ldots(14)$$

and on this convention a positive value of e means that the ray of parameter $\pm b$ *diverges* from the principal ray. Of course the angular aberration in the imaging of a point object at a distance u is simply related to e through the equation

$$\Delta\theta = e - \tfrac{1}{2}\frac{b^3}{f_0 uv}. \qquad \ldots\ldots(15)$$

§ 3. PROPERTIES OF THE HOMOGENEOUS PRISM

Consider a field of strength, H, constant over a circular section of radius a and zero outside this circle. The deviation $2.\phi_b$ of a ray of parameters b and $(H\rho)$ passing through this field is given by

$$\cot\phi_b = (\rho + b)/a\cos\lambda, \qquad \ldots\ldots(16)$$

where $\sin\lambda = b/a$, and ρ is the constant radius of curvature of the track in the field, and is given by $(H\rho)/H$. The focal length of the prism is found by combining (9) and (16). The limiting focal length for infinitely small apertures $\pm b$ is given by $f_0 = (a^2 + \rho^2)/2a$. The general value of f_b can of course be expressed from these two equations but the algebraic form is rather clumsy. As ϕ_b depends only on the rations b/a and ρ/a it is sufficient to consider a single case $a = 10$ cm. and to describe the optical properties for this constant radius with different values of ρ. In figure 4 values of $f_b - f_0$ are plotted for different values of ρ and b, thus giving directly the longitudinal focusing in an initially parallel bundle. For the particular case of $\rho = a$, $\cot\phi_0 = 1$, so that the principal ray is deviated through an angle $\pi/2$. Further, $\cot\phi_b = \sqrt{(a-b)/(a+b)}$, so that f_b becomes constant for all values of $b\,(-a \leqslant b \leqslant +a)$ and takes the value a in agreement with the general expressing for f_0 given above. Thus the aberration completely dissappears over the whole aperture, a result that Korsunsky (1945) has used as the basis of a proposed β-spectroscope.

It is seen from the graphs first, that $f_b - f_0$ steadily increases at first with $\rho - a$, and secondly that the sign of $(f_b - f_0)$ for a given value of ρ changes with the sign of b. The position of minimum confusion lies, therefore, close to the paraxial focus and the spread of the image in this position is terminated sharply by the principal ray. This sharp edge is similar to that associated with images obtained by the method of semicircular focusing. As the sign of the aberration changes with that of $(\rho - a)$, this sharp edge lies on the side of the image spread that is closest to the principal ray belonging to the value of ρ equal to a.

The graphs show that for values of ρ that are not too small, $f_b - f_0$ is roughly proportional to b, so that image formation of a point not at infinity can, according to (13), only be partly corrected, and that for only one side of the aperture. For values of $(\rho - a)/a$ between 0 and 0·05 approximately, this partial correction occurs at object distances with which the associated magnification varies from infinity to unity respectively. For values of $(\rho - a)/a$ greater than 0·05 the correction fails owing to the relative smallness of the quadratic term.

It may appear somewhat surprising that the optical performance of the prism deteriorates as ρ increases and therefore as the deviation of the rays decreases. That this is so can be seen by considering the angular aberration of the focusing of a parallel bundle, which can be found from (14). For very large values of ρ, e becomes $b^2/a\rho$, while the corresponding value of the deviation of a principal ray $2.\phi_0$ becomes $2.a/\rho$. Now the badness in the focusing performance can be judged by some simple multiple of this angular aberration relative to the deviation of the principal ray, say $1000e/2\phi_0$, and as ρ gets very large this becomes $500b^2/a^2$. Table 1 gives the absolute values of $1000e/2\phi_0$ for different values of ρ. As e is not quite symmetrical for positive and negative values of b, the average of the absolute numerical values for $b = \pm 1$ are given.

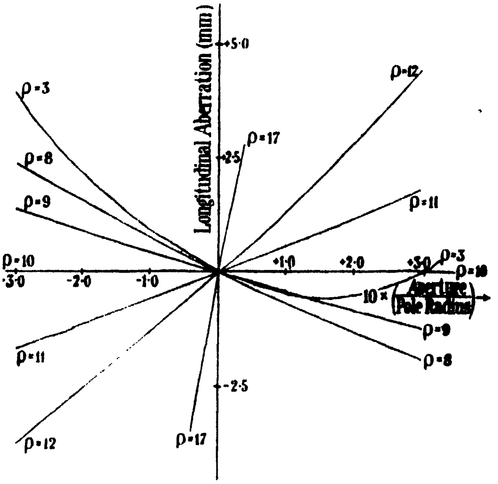


Figure 4. Homogeneous field radius 10 cm., longitudinal aberration ($f_b - f_0$)

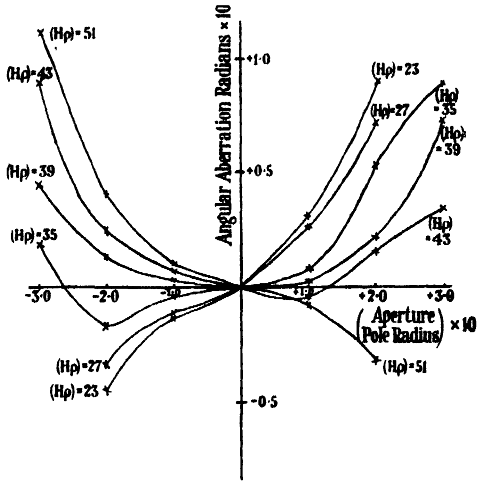


Figure 5.

Table 1

ρ (cm.)	3	8	9	10	11	12	17	50	∞
f_0 (cm.)	5·45	8·20	9·05	10	11·05	12·20	19·45	130·00	∞
Average relative angular aberration $1000\,e/2\phi_0$ ..	0·90	0·61	0·31	0	0·32	0·67	2·00	4·56	5·00

§ 4 PRISM OF TWO POLES SEPARATED A DISTANCE EQUAL TO THEIR COMMON RADIUS

In this case θ_b can be obtained by computing in principle the path of the ray of given parameters b and $(H\rho)$ through the field of the prism. As before, a fixed pole-radius of 10 cm. was taken and the axial component of the field and its axial derivatives were obtained from tables prepared by Dr. L. J. Comrie.* The tables extend only as far as $r = 32$ cm., but for values of r greater than this, H_z and its integral with respect to r can be found from the first few terms of the well known expansion in terms of Legendre polynomials.

The method of computing was to evaluate $\sin\psi$ along a trajectory using equation (8):—

$$r\sin\psi - b = \frac{1}{(H\rho)}\int_{\infty}^{r} rH_z\,dr.$$

The value of $\sin\psi$ given by this equation increases as r gets smaller and finally becomes greater than unity. By inverse interpolation for $\sin\psi = 1$ the value of $r\psi_{\text{min}}$ (the closest real approach of the electron to the axis) is found. The value of θ at this point is the θ_b used above, and is the only value finally required. θ as a function of r is found from

$$\frac{d\theta}{dr} = \frac{1}{r}\tan\psi \qquad \text{......(17)}$$

by numerical integration. Thus the value of θ_b was found by directly integrating this equation up to the value $r = r\psi_{\text{min}}$ without determining intermediate values. There is a singularity in the neighbourhood of $r\psi_{\text{min}}$, as ψ here approaches $\pi/2$. The integrations were performed by the Scientific Computing Service Ltd., and Dr. H. O. Hartley who directed this work overcame the difficulty by transforming the variable of integration to $\sqrt{r - r_{\text{min}}}$. Values of θ_b for some values of the parameters b and $(H\rho)$ are given in table 2. If this table is used with expression (9) it gives all information within its range required concerning the optical properties of the prism. The limiting focal length at infinitely small apertures f_0 is simply given by $\frac{1}{2}\,.\,db/d\theta$ at $b = 0$, which can be computed by the usual methods after differencing table 2. It is quite sufficient to average the first differences, as the precise value of f_0 is of little physical importance.

Table 2

b $(H\rho)\to$	23	27	35	39	43	51
−3			0·7276	0·7697	0·8119	0·8872
−2	0·6693	0·7030	0·7767	0·8157	0·8527	0·9213
−1	0·7291	0·7601	0·8272	0·8620	0·8952	0·9569
0	0·7673	0·8160	0·8768	0·9082	0·9379	0·9934
+1	0·8434	0·8705	0·9258	0·9545	0·9812	1·0308
+2	0·8985	0·9237	0·9734	1·0003	1·0236	1·0692
+3			1·0225	1·0453	1·0667	1·1076

By examining the differences of table 2 it is found that the best focus occurs for values of $(H\rho)$ in the neighbourhood of 39, corresponding to a deviation of a principal ray through $2\phi_0 = \pi - 2\theta_0 = 1{\cdot}3252$ radians. If f_0 is taken as 10·82 cm. for this value of $(H\rho)$, the angular aberrations take the values δf given in the second row of table 3. They are clearly not symmetrical in b. One reason for such asymmetry would lie in an incorrect choice of f_0. Corresponding to a slight change of the position of focus δf a small quantity $\alpha|b|$ where $\alpha = \delta f/f_0^2$ must be added to the angular aberration. If the value of α is chosen as shown in the third row of table 3,

---

* To be published.

the new angular aberration becomes within the computing error proportional to b^3, the coefficient of proportionality being slightly different according as b is positive or negative. The last two rows illustrate this proportionality.

Table 3

b	-3	-2	-1	0	$+1$	$+2$	$+3$
e	0·0039	0·0009	0·0001	0	−0·0001	0·0017	0·0067
ab	0·0006	0·0004	0·0002	0	0·0002	0·0004	0·0006
$e+ab$	0·0045	0·0013	0·0003	0	0·0001	0·0021	0·0073
$0{\cdot}00016b^3$	0·0043	0·0013	0·00016	0			
$0{\cdot}00026b^3$				0	0·00026	0·0021	0·0072

Two features of the aberrations e are of great importance. First they are of correct sign for both positive and negative values of b for correction according to (12) and they have the correct functional variation with b. The maximum value of the second term in (12) is $\frac{1}{2}(b/f_0)^3 \sim 0{\cdot}0001b^3$. Comparison of this with the last two rows of table 3 shows that it is not quite large enough for complete correction, but that it has the right sign for correction all over the aperture.

The second important feature of these aberrations is that they are of the right sign of correction with magnetic lenses, and in magnitude they are as small as, or smaller than, those of the very highest quality magnetic lenses envisaged by Rebsch (1938) in his considerations on the limit of resolution of the electron microscope.

For values of the deviation which become greater or less than the value considered above, the size of the aberration steadily increases and shows more and more marked asymmetry for positive and negative values of b. In figure 5 values of the angular aberration e derived from table 2 are plotted. Inspection of these curves shows that the asymmetry which is already small for $(H\rho)=39$ would be completely removed for the approximate value 40 of this parameter. The angular aberration would then follow a cubic law of variation with b, and the size of the coefficient C in the relation $e=C(b/f)^3$ would take the value $0{\cdot}0002f_0^3 \sim 0{\cdot}25$ for both positive and negative values of b.

§ 5. PRISMS WITH DIFFERENT POLE SEPARATIONS AND FIELD SHAPES

Strictly, the properties of any field can be determined only by ray tracing in the field by some such method as that described above. It is of interest, however, to speculate how the focusing properties change as the field distribution changes by altering the pole separation, or through other causes such as general leakage from the yoke of the magnet.

For the homogeneous prism discussed above there is a complete absence of aberration in the parallel beam formed by a point object at $u=f_0=a$, over all positive and negative values of b. When the poles are separated by a distance equal to their common radius, the best image occurs for an object distance of $2f_0$, but the correcting action of the term containing uv is slightly too small. If the former prism is regarded as the limiting case of a two-pole prism with infinitely small pole separation, it may be expected that the optical properties of this class of prism would show a gradual transition as the pole separation increases. In this case, for each pole separation up to a value somewhat less than the radius of the pole, there will be one position of the object that gives a corrected conjugate image for positive and negative values of b. Each object position, corresponding to each pole separation, will only give rise to a fully corrected image for a particular value of $(H\rho)$. The deviations of the principal rays associated with these values of $(H\rho)$ will decrease slightly as the pole separation increases, the maximum deviation being $\pi/2$ for zero pole separation.

Concerning other alterations of the field distribution, little can be said except that their effect will be less than that which occurs in the corresponding problem associated with the optical properties of magnetic prisms. For the rate of deviation of a ray in the case of a prism depends directly on the field strength, whereas for the prism it depends on the axial derivative of the field.

§ 6. REAL AND VIRTUAL IMAGES

The above considerations on image formation by magnetic prisms has been based on a study of the straight rays that occur where the field strength is negligibly strong. The images are therefore in general virtual. Again, a real object placed where the field strength cannot be neglected, will have associated with it a virtual object, and it is such objects that the method used here considers. Thus this paper considers only those objects and images which are produced or observed by an external lens system, or which lie where the field strength is negligible. The aberration of a real point is not simply derivable from that of the associated virtual point unless they occur where the field is weak though not negligible. Suppose that in figure 6 P is a virtual object-point at a distance r_0 from the axis of the field. The real image associated with it will be free from aberration only if the actual trajectories of the rays belonging to it have a common solution $(r\ , \theta_1)$ for all values of the parameter b. If b is written approximately as $r_0\theta_0$, this means that $(\theta - \theta_0)$ must be constant for all values of θ_0 and one value of r. Provided that ψ is small (8) and (17) can be written

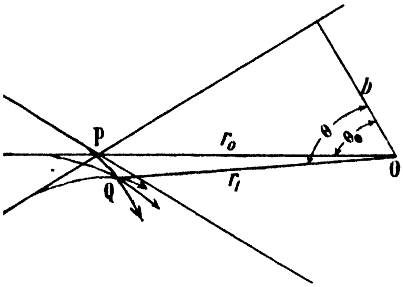


Figure 6. Real and virtual points.

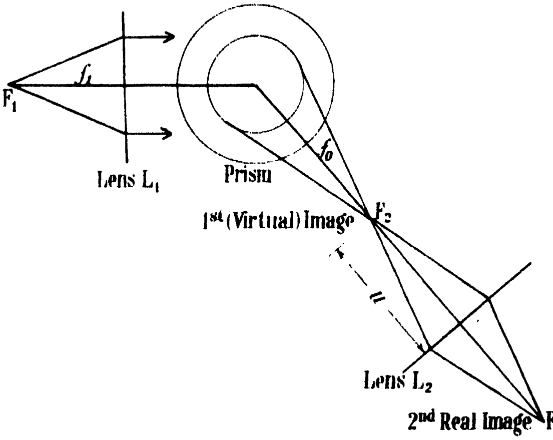


Figure 7. Prism spectrometer with two lenses.

$$-d\theta = \frac{1}{r}\left\{b + \frac{1}{(H\rho)}\int_{\infty}^{r} rH_z\,dr\right\},$$

if θ is taken increasing as r decreases according to figure 6. Thus

$$\theta = b/r + f(r),$$

where the function f is of constant form characteristic of the field distribution. Thus

$$(\theta - \theta_0) = b/r - b/r_0 + f(r).$$

This can only be constant when $r = r_0$. So that provided that $\tan\psi = \sin\psi$, the real and virtual points have the same quality but are separated by a distance $r_0 \,.\, f(r_0)$. In general these conditions are satisfied for both object and image points only when the magnification is in the neighbourhood of unity, and then only when the pole separation is not too large. The general problem of the quality of real image formation could be solved by the method of this paper provided that the integration of equation (17) was carried out step by step and the rays were fully plotted. This has not been attempted here.

§ 7. THE DESIGN OF PRISM β-SPECTROSCOPES

The overall design of a prism β-spectroscope can be made only when a complete knowledge of the focusing properties of prisms with respect to rays not confined to the plane of symmetry is available to supplement the study of the rays in that plane made here. Rays in the plane of the axis, for example, but not in the plane of symmetry, will be focused towards the latter, but in general this focus will be astigmatic with respect to the focus of the rays confined in that plane. Subject to these limitations a preliminary discussion of the merits of some different possible types of prism β-spectrometers can be given as follows. The prism spectrometers discussed are composed of (1) a single prism, (2) a combination of prisms, (3) a combination of a prism with axially symmetric lenses.

(1) *The single prism*

Coggershall and Muskat (1944) have considered the use of a single prism under conditions of limitingly small apertures which they effect by using a stop placed in the magnetic field in the neighbourhood of the turning points of the rays at $r_{\min}$, the rays originating from a point source. Such a design is intended for mass spectroscopy. For β-spectroscopy wide-angle focusing is desired. High resolving power will be obtained under conditions of highly corrected image formation, the conditions for which were discussed for the case of a pair of perfect poles in §§ 4 and 5. Since the use of a single prism requires the observation of real images, the restrictions of § 6 must be satisfied. As the pair of poles are separated farther and farther the magnetic field spreads out more and more beyond the circumference of the circle of the pole $r = a$. At the same time the distance that the object has to be placed from the axis in order to obtain a corrected image increases, so that for each pole separation, a real object at the position of minimum aberration will give rise to an image of much the same quality as a virtual object would if it was placed at the same distance from the axis. In particular, corrected image formation can be achieved by working at unit magnification with a pole separation somewhat less than the pole radius. For a particular value of $(H\rho)$ corresponding to a value somewhat greater than $2.\phi_0 = 1{\cdot}3252$, for the deviation of a principal ray, the spectrometer will have a very high resolving power. The resolving power will steadily decrease as the deviation increases or decreases from this optimum value.

(2) *A combination of prisms*

Korsunsky (1945) has suggested the use of two prisms in series with the object at the focal point of the first, and the image at the focal point of the second. This appears to have no advantage over the use of a single prism. Under the conditions proposed by him for its use, the angular aberrations in the first prism are added to those produced by the second. Thus the increase of dispersion over that due to a single prism is not accompanied by a corresponding increase of resolving power. The aberrations produced in one prism could be made to correct those produced in the other if the direction of deviation was chosen for the second prism so that the parameters b for any ray passing through the combination took different signs in the two prisms. But under these conditions the image formation becomes nearly achromatic, so that the dispersion and resolving power disappear. Korsunsky's four-pole spectrometer was developed from a consideration of rays in the plane of symmetry taking the case of zero pole separation. On these grounds advantage from the use of two prisms would only accrue if the quality of image formation for small pole separations and high magnification was superior to that at unit magnification and much larger pole separation. The total focusing of a cone of rays by each prism will be astigmatic and the possibility does remain of using a system of prisms to produce a pair of stigmatic points.

(3) *Combination of a prism with lenses*

Klemperer (1935) described a spectrometer that he constructed and that was modelled roughly on the lines of an optical prism spectrometer. Unfortunately this work had to be abandoned before the best working conditions were found. Now that the focusing action of the prism itself are more clearly realized, an instrument of this type could be designed as shown in figure 7. Here the magnetic lens L_1 forms a parallel bundle which is deviated by the prism and brought to a focus at F_2. In general this image will be virtual and must be observed with another lens L_2. Under certain conditions the aberrations of the prism have opposite sign to those of the lenses, so that corrected image formation at F_3 would be possible over a wide aperture for one value of $(H\rho)$ provided that lenses of sufficiently high quality were available. Unfortunately there is considerable uncertainty in this matter. It is known that the angular aberration introduced by a magnetic lens is proportional to $(b/f)^3$, where b is the intercept of the ray in the plane of the lens, regarded as thin, although it is not clear that this functional variation will hold all over the large apertures considered here. If C_1 is the coefficient of proportionality, so that for a lens the angular aberration becomes $C\ (b/f)^3$, and if C_2 is the corresponding coefficient for the prism, the total angular aberration of the final image of figure 7 is $C\ (b/f_1)^3 - C_2(b/f_0)^3 + C_1(b'/f_2)^3$ where $b' = b(u/f)^3$ and f is the focal length of the first lens, f_0 that of the prism and f_2 that of the second lens. Thus for corrected image formation:—

$$C_1/f_1^3 - C_2/f_0^3 + C_1/(f_0 f_2/u)^3 = 0.$$

For a real final image, u must be greater than f_2 so that for this equation to hold C_2 must be greater than C_1. From the discussion given in §4 it would appear that values of the order of $0{\cdot}25$ can be attained for C_2. For the type of lenses used in β-ray spectroscopy the values of C_2 are not available. From simple theoretical considerations the writer (1942) showed that for a class of short solenoids advantage was to be expected by a reduction of the ratio (diameter/length) which reduction fortunately also favours economy of power. However, the C_1 values have not yet been worked out. In connexion with the electron microscope, Rebsch suggested a lower limit of $C_1=0{\cdot}25$, but his argument was based on considerations of instrumentation that do not necessarily apply to the lenses required for β-spectroscopy. More recently slightly smaller values have been suggested.

Compared with the single prism, the compound spectrometer employing lenses suffers from two disadvantages, in that it is more complicated and that the correction is a much more formidable task. The final choice between the two types depends ultimately on the C values attainable in magnetic lenses, and the astigmatic focusing properties of the prism.

ACKNOWLEDGMENTS

I should like to thank Prof. J. D. Bernal, F.R.S., for his encouragement and generous support. My thanks are due also to Dr. L. Jánossy for discussion concerning the method of ray tracing, to Drs. Comrie and Hartley, of the Scientific Computing Service Ltd., for discussion concerning the numerical integrations, and to Dr. Ehrenberg for general discussions.

REFERENCES

Coggershall and Muskat, 1944. *Phys. Rev.*, **66**, 189.
Herzog, 1934. *Z. Phys.*, **89**, 447.
Klemperer, 1935. *Phil. Mag.*, **20**, 545.
Korsunsky, 1945. *J. Phys. U.S.S.R.*, **9**, 14.
Rebsch, 1938. *Ann. Phys., Lpz.*, **31**, 551.
Sawyer, 1936. *Proc. Camb. Phil. Soc.*, **32**, 453.
Siday, 1942. *Proc. Phys. Soc.*, **54**, 226.

THE THEORY OF RADIATION DAMPING

By J. HAMILTON,
University of Manchester

MS. received 3 *March* 1947

§ 1. INTRODUCTION AND SUMMARY

THIS paper is the result of an attempt to get a better understanding of the quantum theory of radiation damping developed by Weisskopf and Wigner (1930), Heitler (1941), Wilson (1941) and Heitler and Peng (1942). It has been suggested by Peng (1944) that the Heitler-Peng damping equation arises because of the degeneracy of the unperturbed states of the system of particles and radiation; and he has claimed (1946) that a careful mathematical treatment of the "quasi-degenerate" continuum of energy levels not only gives the damping equation, but also removes the well known divergence difficuties of radiation theory.

It is obvious that the concept of degeneracy is much simpler when the system is enclosed in a box, and as a consequence, the energy levels are discrete. The difficulties of the solution then lie rather in the complexity of the mathematics than in the concepts. It has, however, been possible, subject to certain limitations, to solve the emission and scattering problems exactly for a cubic box in the discrete energy case. These limitations are just those imposed in the original Dirac and Heitler-Peng methods of calculating the transition probabilities—namely the omission of states which seem to be physically unimportant. That this limitation is a very severe one, can be seen from the one case in which it has been possible to consider more than the conventional states. The addition of some physically unimportant states to the equations of the emission problem reduces the divergence difficulties. In general, however, it seems impossible to use the discrete energy method to investigate whether or not the divergencies in radiation theory are purely mathematical difficulties due to the neglect of higher-order processes. It is as well, therefore, to make it clear at the outset that it is really models of the actual systems which are considered. For example, the scattering problem corresponds to a large number of oscillators, whose energies are close together, which are coupled to each other through certain intermediate oscillators whose energies are very different from those of the main set.

The questions which then arise are (i) what rôle degeneracy plays in the solution, (ii) why the perturbation method does not give the correct transition probabilities in scattering problems, (iii) how the behaviour of such a model differs from the actual radiation system as observed. Degeneracy, in the strict sense, plays no rôle whatever, as the enclosing box can be chosen so that the ratios of the squares of the edges are irrational numbers. In that case there is no degeneracy. The perturbation method, however, breaks down for strong interactions in scattering problems. A rough criterion can be given when the compound interaction matrix elements * are not strongly dependent on the directions of the incident and scattered particles. Then the product of the average magnitude of the matrix elements with the density function † $\rho(E)$ of the scattered states must be much less than unity ($\hbar=1$) for the perturbation procedure to be applicable. What is happening can easily be seen when the energy levels of the unperturbed system are highly degenerate. The interaction causes a splitting of each energy level, and it is only when the splitting corresponds to a change in energy which is small compared with the energy difference between two adjacent levels that the perturbation method is applicable. For a sufficiently strong interaction, the first-order energy perturbation can be made very large; but the exact solution shows that the levels which split off from a degenerate level E_2 cannot go further than roughly half-way between this degenerate level and the adjacent degenerate levels E_1 and E_3. An exact criterion for the validity of the perturbation method is that the magnitudes of all the eigenvalues of a homogeneous integral equation must be much greater than unity. The kernel of this homogeneous integral equation is the kernel of the Heitler-Peng integral equation; and the two equations are closely related.

The chief difference between the behaviour of the model and the observed

---

* In the normal sense.

† Which also has to be little dependent on the angles.

behaviour of the atomic emitter is the line shift. The final states of the model assume the usual energy distribution or line form; but the centre of the line does not coincide with the initial energy. The shift is very large or infinite. However, the energy *is* conserved. The reason for this strange situation can be seen as follows. The usual line form centred about the resonant frequency gives a small probability for the emission of very high-energy quanta. When the total emitted energy is calculated, it is seen to diverge due to these high-energy quanta. The shift in the line compensates for this effect. The shift is very much reduced if a set of states, which are of no physical importance, is introduced. The usual states are (i) the atom in its excited state, (ii) the atom in its normal state and one photon present; and in the new states (iii) the atom is in its excited state and two photons are present. The very improbable transitions from (ii) to (iii) very largely counteract the shift. In the scattering case the comparison with reality is more complicated. Several terms of the first-order self-energy type $\sum_r |H_r|^2/(E_0-E_r)$ involving both the simple and the compound matrix elements, have to be neglected. When this is done the transition probabilities are identical with the Heitler-Peng result.

It is assumed both in the emission and the scattering cases that terms of the type $\sum_r |H_r|^2/(E_0-E_r)^2$ lead to no difficulties. For this to be true it is only necessary that the contribution to the sum from the region of large (E_r-E_0) is finite. Even when the number of oscillators tends to infinity the sum remains finite, provided a cut-off is made at any energy, however high. (For example, an electron whose radius is 10^{-100} cm. would suffice.)

§2. THE EIGENVALUE PROBLEM

The motion of a number of coupled oscillators can be treated by the transition probability or by the eigenvalue method. The latter method is used here.

Suppose that on neglecting the coupling, the N component parts of the system are described by a Hamiltonian H°, which leads to the N stationary states described by the wave functions $\psi_1, \psi_2, \ldots . \psi_N$; whose energies are $E_1, E_2, \ldots . E_N$ respectively. These wave functions can be written as

$$\psi_r = \phi_r e^{-iE_r t} \qquad (r=1, 2, \ldots . N),$$

where the function ϕ_r is independent of time (and $\hbar=1$). Any state of the system is described by

$$\Psi = \sum_{r=1}^{N} a_r \psi_r = \sum_{r=1}^{N} a_r \phi_r e^{-iE_r t}, \qquad \ldots\ldots(1)$$

the a_r being constants when there is no coupling.

Introducing a time-independent interaction V between the components of the system, the equation of motion becomes

$$i\frac{\partial\Psi}{\partial t} = (H^\circ + V)\Psi. \qquad \ldots\ldots(2)$$

If the ϕ_r are chosen as a normal orthogonal set such that*

$$(\overline{\phi_r} \,.\, \phi_s)_v = \delta_{rs},$$

---

* $(\ldots\ldots)_v$ represents the integral over the volume in which the system is contained.

then (2) gives

$$i\frac{d}{dt}(a_r e^{-iE_r t}) = \sum_{s=1}^{N} H_{rs} a_s e^{-iE_s t}, \qquad \ldots\ldots(3)$$

where $H_{rs} = (\bar{\phi}_r \,.\, H_0 + V \,.\, \phi_s)_v$, and is independent of the time.

The set (3) can be solved in terms of a series of N "normal modes" of amplitudes A_r and energies Λ by substituting

$$a_r e^{-iE_r t} = A_r e^{-i\Lambda t}.$$

This gives

$$A_r \,.\, \Lambda = \sum_{s=1}^{N} H_{rs} A_s, \qquad \ldots\ldots(4)$$

and hence the Λ are the roots of

$$|H_{rs} - \Lambda \,.\, I| = 0. \qquad \ldots\ldots(5)$$

H_{rs} being Hermitian, the N roots Λ_μ ($\mu = 1, 2, \ldots. N$) are real; and denoting the amplitudes of the μth mode by A_r^μ, the normal modes are

$$\Psi'^{\mu} = \sum_{r=1}^{N} a_r^\mu \psi_r, \quad \text{where} \quad a_r^\mu = A_r^\mu \,.\, e^{-i(\Lambda_\mu - E_r)t}.$$

The A_r^μ can satisfy the conditions

$$\left.\begin{aligned} &\sum_{r=1}^{N} \bar{A}_r^\nu A_r^\mu = \delta_{\mu\nu} \qquad (\mu, \nu = 1, 2, \ldots. N), \\ &\text{and it follows that} \\ &\sum_{\nu=1}^{N} \bar{A}_r^\nu A_s^\nu = \delta_{rs} \qquad (r, s = 1, 2, \ldots. N). \end{aligned}\right\} \ldots\ldots(6)$$

The general state of the coupled system can be written

$$\Psi' = \sum_{\mu=1}^{N} c_\mu \Psi'^{\mu}, \quad \text{where} \quad \sum_{\mu=1}^{N} |c_\mu|^2 = 1;$$

so from (1)

$$a_r e^{-iE_r t} = \sum_{\nu=1}^{N} c_\nu A_r^\nu e^{-i\Lambda_\nu t}.$$

If the initial conditions are $a_r(0)$ ($r = 1, 2, \ldots. N$), then

$$a_r(0) = \sum_{\nu=1}^{N} c_\nu A_r^\nu$$

or

$$c_\nu = \sum_{r=1}^{N} \bar{A}_r^\nu a_r(0) \qquad (\nu = 1, 2, \ldots. N). \qquad \ldots\ldots(7)$$

The A_r^ν are merely the coefficients of the unitary matrix S which transforms the system from the representation in which H° is diagonal to that in which $H^\circ + V$ is diagonal, viz.: $S_{ik} = A_i^k$.

The conservation and spread of the energy of the oscillators due to the interaction can be readily deduced. The average energy $\bar{E}$ of a state $\overline{\Psi'}$ is

$$\bar{E} = \left(\overline{\Psi'}, i\frac{\partial}{\partial t}\Psi'\right)_v.$$

This gives

$$\bar{E} = \sum_{\mu=1}^{N} |c_\mu|^2 \Lambda_\mu,$$

and similarly

$$\overline{E^2} = \sum_{\mu=1}^{N} |c_\mu|^2 \Lambda_\mu^2.$$

If the initial conditions are

$$\left.\begin{aligned} a_k(0) &= 1, \\ a_r(0) &= 0, \quad r \neq k, \end{aligned}\right\} \qquad \ldots\ldots(8)$$

then $c_\nu = \bar{A}_k^\nu$, so

$$\bar{E} = \sum_{\mu=1}^{N} |A_k^\mu|^2 . \Lambda_\mu = \sum_{\mu=1}^{N} \overline{A_k^\mu} \sum_{s=1}^{N} H_{ks} A_s^\mu = \sum_{s=1}^{N} H_{ks} \delta_{ks}.$$

Thus

$$\bar{E} = E_k + V_{kk}. \qquad \ldots\ldots(9)$$

Similarly

$$\overline{E^2} = \sum_{r=1}^{N} \overline{|H_{kr}|^2},$$

and

$$\overline{(E - E_k)^2} = \sum_{r=1}^{N} |V_{kr}|^2. \qquad \ldots\ldots(10)$$

§ 3. THE FUNDAMENTAL EQUATIONS

(a) *Atomic emission*

The system consists of an atom enclosed in a box with perfectly reflecting walls. Following the usual treatment, the states considered are:—

ψ_0—the atom in its excited state of energy E_0 with no photons present.

ψ_r—the atom in its normal state of energy E, and a photon of energy ϵ_r present.

It is assumed that there are N states of the latter kind. The interaction V allows single quantum jumps only, i.e. $V_{rs} = 0$ unless one of r or s is zero. Putting $V_{0r} = H_{0r}$, the fundamental equations (4) connecting the $N+1$ states ψ_i ($i = 0$, $1, \ldots N$) become

$$\left.\begin{aligned} &(E_0 - \Lambda) A_0 + \sum_{s=1}^{N} H_{0s} A_s = 0, \\ &(E + \epsilon_r - \Lambda) A_r + H_{r0} A_0 = 0 \qquad (r = 1, 2, \ldots . N). \end{aligned}\right\} \ldots\ldots(11)$$

If two or more energies ϵ_r are equal, degenerate solutions will exist. Let $\epsilon_{r_1} = \epsilon_{r_2} = \ldots = \epsilon_{r_m}$ be degenerate levels. Then the degenerate solution satisfies

$$\left.\begin{aligned} &H_{r_i} A_0 = 0 \qquad (i = 1, 2, \ldots m), \text{ i.e. } A_0 = 0, \\ &A_s = 0 \qquad \epsilon_s \neq \epsilon_{r_i}, \\ &\sum_{i=1}^{mr} H_{0r_i} A_{r_i} = 0. \end{aligned}\right\} \ldots\ldots(12)$$

The last conditions shows that the vector $(A_{r_{11}}, \ldots A_{r_m})$ must be orthogonal to the vector $(\bar{H}_{0r_1}, \ldots \bar{H}_{0r_m})$, so there are $(m-1)$ independent degenerate solutions.

If r denotes the various degenerate levels, the level ϵ_r being m_r-fold and Λ_μ ($\mu=0$, 1, ... M) the remaining roots such that $\Lambda_\mu \neq E+\epsilon_s$ for any s, it might be thought that the general solution of equation (3) should be

$$a_k e^{-iE_k t} = \sum_{\mu=0}^{M} A_k^{\mu} e^{-i\Lambda_\mu t} + \sum_{r} \sum_{\beta=1}^{m_r-1} A_k^{\beta} \left(\frac{t}{i}\right)^{\beta-1} e^{-i(E+\epsilon_r)t}.$$

On substituting this expression in (3) and using relations (12), it can be seen that the A_k^{β} vanish except when $\beta=1$, so no powers of t occur in the solution. For the emission problem the initial conditions are

$$\left.\begin{array}{ll} a_0(0)=1, & \\ a_r(0)=0 & (r=1, 2, \ldots N). \end{array}\right\} \qquad \ldots\ldots(13)$$

Then $c_\mu = \tilde{A}_0^{\mu}$; and if $\Lambda_\nu = E+\epsilon_r$ then $c_\nu = 0$. So such degenerate states do not enter into the solution.

To find the non-degenerate solutions, multiply the first equation (11) by $\prod_{r=1}^{N} (E+\epsilon_r-\Lambda)$, and substitute for A_s, using the second equation (11), giving

$$\left\{(E_0-\Lambda) \prod_{r=1}^{N} (E+\epsilon_r-\Lambda) - \sum_{s=1}^{N} \prod_{(s)\, r=1}^{N} (E+\epsilon_r-\Lambda)\, |H_{0s}|^2\right\} A_0 = 0\,{}^*$$

Thus if $A_0 \neq 0$, and if there are m_r states with energy $E+\epsilon_r$, division by

$$\prod_{r=1}^{N} (E+\epsilon_r-\Lambda)^{m_r-1}$$

gives

$$\Lambda - E_0 + \sum_{r=1}^{M} \left\{ \sum_{i=1}^{m_r} |H_{0r_i}|^2/(E+\epsilon_r-\Lambda) \right\} = 0, \qquad \ldots\ldots(14)$$

where M is the number of distinct ϵ_r values and r_i ($i=1, 2, \ldots m_r$) refer to the states with energy $E+\epsilon_r$.

The function

$$f(\Lambda) = \Lambda - E_0 + \sum_{r=1}^{M} \left\{ \sum_{i=1}^{m_r} |H_{0r_i}|^2/(E+\epsilon_r-\Lambda) \right\}$$

has simple poles at $\Lambda = E+\epsilon_r$, and its derivative

$$\frac{df(\Lambda)}{d\Lambda} = 1 + \sum_{r=1}^{M} \left\{ \sum_{i=1}^{m_r} |H_{0r_i}|^2/(E+\epsilon_r-\Lambda)^2 \right\}$$

has poles of order 2 at $\Lambda = E+\epsilon_r$, and for real $\Lambda \frac{df(\Lambda)}{d\Lambda} > 1$. Hence the equation $f(\Lambda)=0$ has one solution between $E+\epsilon_r$ and $E+\epsilon_{r+1}$ ($r=1, 2, \ldots M-1$) and the 1st and $(M+1)$th solutions are less than $E+\epsilon_0$, and greater than $E+\epsilon_M$, respectively, these being the least and the greatest energy levels†. Figure 1 shows the graphical solution.

If $\sum_{s=1}^{N} |H_{0s}|^2/(E+\epsilon_s) \to \infty$ as $N \to \infty$, the lowest root $\Lambda \to -\infty$, and the root between $E+\epsilon_r$ and $E+\epsilon_{r+1}$ tends to $E+\epsilon_r$ as the limit. Thus N must in general be kept finite until the solution is obtained.

---

* $\prod_{(s)\, r=1}^{N} = \prod_{r=1}^{s-1} \cdot \prod_{r=s+1}^{N}$.

† A footnote to Weisskopf and Wigner's (1930) paper shows that they were aware of equation (14) and its chief properties.

(b) *Scattering*

The scattering of a photon or a meson by electrons or nucleons is considered. The system consists of the scattering particle enclosed in a box so that a large number of different photon or meson states can occur. Initially there is one photon or meson, and only those scattered states in which one quantum is present are allowed. Intermediate states can occur, but it is necessary to impose the condition that the energy of any intermediate state is far removed from the energies of the initial and the final states—a physically reasonable assumption. The interaction can lead to either one or two quantum jumps.

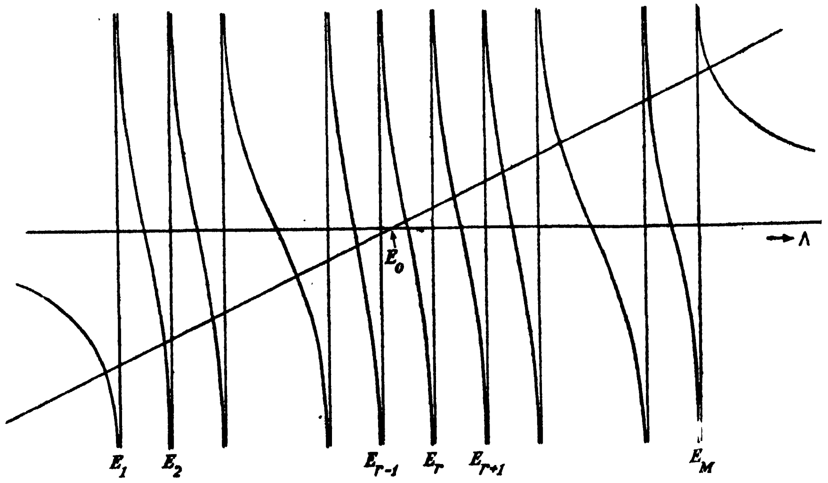


Figure 1. The intersections of the functions $\Lambda - E_0$ and $\sum_{r=1}^{M} |H_r|^2/(\Lambda - E_r)$.

Denoting the initial, intermediate and final states by the subscript indices 0; i, j; r, s; the fundamental equations (4) become

$$\left.\begin{aligned}&(E_0-\Lambda)A_0+\sum_i H_{0i}A_i+\sum_s H_{0s}A_s=0,\\&H_{i0}A_0+(E_i-\Lambda)A_i+\sum_s H_{is}A_s=0,\\&H_{s0}A_0+\sum_i H_{si}A_i+(E_s-\Lambda)A_s+\sum_r{}_{(s)}H_{sr}A_r=0^*.\end{aligned}\right\}\quad\ldots\ldots(15)$$

If there are P intermediate and M final states, this set of equations gives $(1+P+M)$ roots Λ on eliminating all the A. It is more convenient to eliminate the A_i first, giving:—

$$\left.\begin{aligned}&\left\{E_0-\Lambda-\sum_i\frac{H_{0i}H_{i0}}{E_i-\Lambda}\right\}A_0+\sum_s\left\{H_{0s}-\sum_i\frac{H_{0i}H_{is}}{E_i-\Lambda}\right\}A_s=0,\\&\left\{H_{s0}-\sum_i\frac{H_{si}H_{i0}}{E_i-\Lambda}\right\}A_0+\left\{E_s-\Lambda-\sum_i\frac{H_{si}H_{is}}{E_i-\Lambda}\right\}A_s\\&\qquad+\sum_r{}_{(s)}\left\{H_{sr}-\sum_i\frac{H_{si}H_{ir}}{E_i-\Lambda}\right\}A_r=0\quad(s=1,2,\ldots M).\end{aligned}\right\}\quad\ldots\ldots(16)$$

---

* $\sum_r{}_{(s)}$ denotes the summation over all final states except s.

The terms $\sum_i \frac{|H_{0i}|^2}{E_i-\Lambda}$, $\sum_i \frac{|H_{si}|^2}{E_i-\Lambda}$ are clearly second-order self-energy terms. Further, the remaining summations containing $E_i-\Lambda$ in the denominator in general only contain a finite number of terms. If only those values of Λ which are in the vicinity of the initial energy are to be considered, Λ can be replaced in these summations by E_0 (as then $|E_i-\Lambda| \gg |E_0-\Lambda|$). Writing

$$E'_s=E_s-\sum_i \frac{|H_{si}|^2}{E_i-\Lambda}, \qquad H'_{sr}=H_{sr}-\sum \frac{H_{si}H_{ir}}{E_i-E_0}, \text{ etc.}$$

the set (16) becomes

$$\left.\begin{aligned} &(E'_0-\Lambda)A_0+\sum_s H'_{0s}A_s=0, \\ &H'_{s0}A_0+(E'_s-\Lambda)A_s+\sum_{r}{}_{(s)}H'_{sr}A_r=0. \ (s=1, 2, \ldots M). \end{aligned}\right\} \quad \ldots\ldots(17)$$

The H'_{sr} are the matrix elements used by Heitler and Peng (1942). The P roots neglected in the transition from (16) to (17) correspond to states in which the amplitudes of the intermediate states are relatively large. Set (17) is a good approximation so long as its $(M+1)$ roots Λ are much closer to E_0 than to any of the E_i. Including 0 in the s, and neglecting the dashes, (17) becomes

$$(E_s-\Lambda)A_s+\sum_{r}{}_{(s)}H_{sr}A_r=0 \qquad (s=0, 1, 2, \ldots M) \qquad \ldots\ldots(18)$$

The solution of this set of equations requires further knowledge about the energy levels of the system.

§ 4. THE ENERGY LEVELS OF THE SYSTEM AND THE PHYSICAL MODEL

(a) *Atomic emission*

The box enclosing the system can be taken as a cube with sides of length L. A radiation oscillator of vector momentum $\boldsymbol{p}$ is described by a wave function

$$\phi_p = L^{-3/2}e^{-i(\boldsymbol{p}.\boldsymbol{x})/\hbar}$$

where $\boldsymbol{x}$ is the position vector. The following periodicity conditions are imposed:

$$p_x \,.\, L=\boldsymbol{h} \,.\, n_x,$$
$$p_y \,.\, L=\boldsymbol{h} \,.\, n_y,$$
$$p_z \,.\, L=\boldsymbol{h} \,.\, n_z,$$

where n_x, n_y, n_z are integers, positive, negative or zero. Thus

$$p=\frac{\boldsymbol{h}}{L}\sqrt{n_x^2+n_y^2+n_z^2}=\frac{\boldsymbol{h}}{L}\sqrt{n}, \text{ say.}$$

For some values of the positive integer n, the equation

$$n_x^2+n^2+n_z^2=n \qquad \ldots\ldots(19)$$

has no solutions, but in general it has many. The average number of sets (n_x, n_y, n_z) which give a value of $\sqrt{n_x^2+n^2+n_z^2}$ lying in the range $(\sqrt{n}, \sqrt{n}+\delta)$ is $4\pi n \,.\, \delta$. For large n,

$$\sqrt{n+1}-\sqrt{n} \simeq 1/2\sqrt{n},$$

so the average number of solutions of (19) for each n is $2\pi\sqrt{n}$. The average number of states with the same momentum p is $2\pi\sqrt{n}=2\pi \,.\, pL/\boldsymbol{h}$, and the difference between adjacent values of p is, approximately, $(\boldsymbol{h}/L) \,.\, 1/2\sqrt{n}=(\boldsymbol{h}/L)^2/2p$, or a small integral multiple thereof.

For photons, the energy ϵ is given by $\epsilon = \mathbf{c}p$, so the average number of states of the same energy ϵ is $2\pi\epsilon \, . \, L/\mathbf{c}h$ and the separation of adjacent energy levels is $(\mathbf{c}h/L)^2/2\epsilon$, or a small integral multiple of it. [The total number of states in the energy range $(\epsilon, \epsilon + d\epsilon)$ is the well known expression $4\pi\epsilon^2 d\epsilon \, . \, (L/\mathbf{c}h)^3$.]

It is clear that the energy levels are mostly highly degenerate, and that the distinct energy levels lie very close together. If the total number of oscillators N is kept finite, the atomic emission problem, developed in § 3 above, can be solved approximately, as it stands. Comparing the series

$$\sum_{\mu=0}^{M} \frac{|A^\mu|^2}{(E+\epsilon_r-\Lambda)^2} \quad \text{and} \quad \sum_{\mu=0}^{M} \frac{1}{(E+\epsilon_r-\Lambda)^2}$$

(which can be evaluated after some elementary but laborious manipulation) it is possible to estimate $|A^\mu_r|^2$, and hence find the complete solution. A simpler method, suitable for both the emission and the scattering problems, will, however, be used below.

(b) *Scattering*

If the momenta of the incident radiation and the scatterer are equal and opposite, the total momentum will always be zero. For simplicity, consider the scattering of a photon by a particle of rest mass $\mu/\mathbf{c}^2$. In a final state in which the photon has momentum p, the total energy is

$$E = k + \sqrt{k^2+\mu^2}, \quad \text{where} \quad k = p\mathbf{c}.$$

Thus $E - \mu^2/E = 2k$, and a small change δk in k leads to a small change δE in E given by

$$(1 + \mu^2/E^2) \, . \, \delta E = 2 \, . \, \delta k.^* \qquad \ldots\ldots(20)$$

So the distance between adjacent energy levels is

$$\Delta E = (\mathbf{c}h/L)^2 \, . \, \frac{1}{k(1+\mu^2/E^2)}$$

or a small integral multiple thereof; and the average degree of degeneracy of each level is†

$$m = 2\pi k \, . \, (L/\mathbf{c}h),$$
$$= \pi(E - \mu^2/E) \, . \, (L/\mathbf{c}h).$$

If the intervals between adjacent energy levels were equal in the vicinity of the initial energy value and if all energy levels in this region had the same degree of degeneracy, the solution for the emission and the scattering cases would be very much simplified. In the emission case the line breadth has a finite value irrespective of the size of the enclosing box, so it is the average behaviour of the A^μ_r over a large number of energy levels which is important. Therefore it is not be expected that the change in the solution due to altering the energy levels to make them equally degenerate will be of any practical importance. The scattering problem will also be treated for such a physical model in which the energy levels are equally spaced, and all are equally degenerate (in the vicinity of the initial

---

* For the scattering of a meson of rest mass $\mu'/\mathbf{c}^2$ by a particle of rest mass $\mu/\mathbf{c}^2$ equation (20) becomes

$$\{E^3 - (\mu^2 - \mu'^2)^2\} \, . \, \delta E = 4kE^3 \, . \, \delta k.$$

† Of course there may be a further degeneracy for given energy and direction due to spin, polarization, etc.

energy). The transition probabilities can only depend on some function of the interaction matrix $\boldsymbol{H}$ and on the density of the final states over a small region which may contain as many energy levels as is desired.

§ 5. THE SOLUTION OF THE EMISSION PROBLEM

It is further assumed that, over the small energy range about the initial energy value E_0, in which the energy levels are taken to be equally spaced and equally degenerate, the matrix elements H are independent of the energy of the photons. The error thus introduced will be seen to be negligible. Besides the equation

$$\Lambda - E_0 + \sum_{r=1}^{M} \left\{ \sum_{i=1}^{mr} |H_{0r_i}|^2/(E+\epsilon_r - \Lambda) \right\} = 0; \qquad \ldots\ldots(14)$$

the normalizing condition for any solution, viz.:—

$$|A^\mu|^2 \left[1 + \sum_{r=1}^{M} \left\{ \sum_{i=1}^{mr} |H_{0r_i}|^2/(E+\epsilon_r - \Lambda)^2 \right\} \right] = 1. \qquad \ldots\ldots(21)$$

is required. This is merely the equation $|A_0^\mu|^2 + \sum_{r=1}^{N} |A_r^\mu|^2 = 1$. Put $E_r = E + \epsilon_r$. The position of the root Λ which lies in the vicinity of some energy value E' is obtained as follows:—

$$\begin{aligned} \Lambda - E_0 &= \sum_{r=1}^{M} \left\{ \sum_{i=1}^{mr} |H_{0r_i}|^2/(\Lambda - E_r) \right\} \\ &= \sum_{r=s-n}^{s+n} \left\{ \sum_{i=1}^{mr} |H_{0r_i}|^2/(\Lambda - E_r) \right\} + \left(\sum_{r=1}^{s-n-1} + \sum_{r=s+n+1}^{M} \right) \left\{ \sum_{i=1}^{mr} |H_{0r_i}|^2/(\Lambda - E_r) \right\}, \end{aligned}$$

where $E' = E_s$, and n is a large positive integer. Further, the summation over the degenerate states of any energy level is equivalent to an angular integration, thus

$$\sum_{i=1}^{mr} |H_{0r_i}|^2 = \frac{m_r}{4\pi} \int |H_{0r_i}|^2 \, d\Omega_i. \qquad \ldots\ldots(22)$$

The first term in the expression above for $\Lambda - E_0$ can be written

$$\sum_{i=1}^{m} |H_{0s_i}|^2 . \sum_{r=-n}^{+n} \frac{1}{(\Lambda - E_s) - r\Delta E},$$

where ΔE is the interval between adjacent energy levels in the vicinity of E_s.

Further,

$$\sum_{r=-\infty}^{+\infty} \frac{1}{(\Lambda - E_s) - r\Delta E} = \lim_{n\to\infty} \sum_{r=-n}^{+n} \frac{1}{(\Lambda - E_s) - r\Delta E} = (\pi/\Delta E) . \cot\{(\Lambda - E_s)\pi/\Delta E\}$$

so

$$\begin{aligned} \Lambda - E_0 = {} & \sum_{i=1}^{m} |H_{0s_i}|^2 . (\pi/\Delta E) . \cot\{(\Lambda - E_s)\pi/\Delta E\} \\ & + \int_{(E_s)} (E_s - E_r)^{-1} . \int |H_{0r}|^2 \, d\Omega . \rho(E_r) dE_r + \eta, \end{aligned} \qquad \ldots\ldots(23)$$

where η is a small error and (E_s) denotes that the range $(E_s - \xi, E_s + \xi)$, where $\xi = n\Delta E$, is excluded from the integration. $\rho(E) . dE . d\Omega$ is the total number of states of energy lying in the range $(E, E + dE)$, and with momenta lying in the angle $d\Omega$. ξ can be given any finite value, e.g. $0{\cdot}00001 \times E'$, and then if the size of the box is sufficiently large the error η can be made as small as desired, compared with the first term on the right-hand side of (23).

Putting

$$R(E_s) = \int_{(E_s)} (E_s - E_r)^{-1} \,.\, \int |H_{0r}|^2 d\Omega \,.\, \rho(E_r) dE_r$$

the equation for Λ becomes

$$\Lambda - E_0 = \pi \,.\, \int |H_{0s}|^2 \, d\Omega \,.\, \rho(E_s) \,.\, \cot\{(\Lambda - E_s)\pi/\Delta E\} + R(E_s) \quad \ldots\ldots(24)$$

Equation (21) is treated similarly, viz.:—

$$\sum_{r=1}^{M} \left\{ \sum_{i=1}^{mr} |H_{0r_i}|^2/(\Lambda - E_r)^2 \right\} = \sum_{i=1}^{m} |\dot{H}_{0s.}|^2 \,.\, \sum_{r=-n}^{+n} \frac{1}{(\Lambda - E_s - r\Delta E)^2}$$
$$+ \int_{(E_s)} (E_s - E)^{-2} \,.\, \int |H_{0r}|^2 d\Omega \,.\, \rho(E) dE,$$

and

$$\sum_{r=-\infty}^{+\infty} \frac{1}{(\Lambda - E_s - r\Delta E)^2} = (\pi/\Delta E)^2 \,.\, \operatorname{cosec}^2\{(\Lambda - E_s)\pi/\Delta E\};$$

so

$$\sum_{r=1}^{M} \left\{ \sum_{i=1}^{mr} |H_{0r_i}|^2/(\Lambda - E_r)^2 \right\} = \pi\rho(E_s) \int |H_{0s}|^2 d\Omega \,.\, (\pi/\Delta E) \operatorname{cosec}^2\{(\Lambda - E_s)\pi/\Delta E\}$$
$$+ \int_{(E_s)} (E_s - E)^{-2} \int |H_{0r}|^2 d\Omega \,.\, \rho(E) dE + \eta'. \quad \ldots\ldots(25)$$

The first term on the right-hand side of (25) increases like $1/\Delta E$ as L increases, and η' the error becomes small compared with the first term for sufficiently large L. The remaining term is independent of L, and, provided the integral is finite, it becomes negligible compared with the first term for sufficiently large L. It is of interest to note that the integral may have any finite value however large. Even for unbound states this could be satisfied by taking a cut-off value (electron radius) of 10^{-100} cm., say.

Thus we have (using (24)

$$|A_0^\mu|^2 \,.\, \left[1 + \pi\rho(E_s) \,.\, \int |H_{0s}|^2 d\Omega \,.\, (\pi/\Delta E) \operatorname{cosec}^2\{(\Lambda - E_s)\pi/\Delta E\} \right] = 1$$

and

$$|A_0^\mu|^2 \,.\, \frac{1}{\Delta E} \to \frac{\rho(E_s) \int |H_{0s}|^2 d\Omega}{\left\{\pi\rho(E_s) \int |H_{0s}|^2 d\Omega\right\}^2 + \{E_s - E_0 - R(E_s)\}^2} \quad \ldots\ldots(26)$$

as $L \to \infty$. $E_s - E_0 - R(E_s)$ is substituted for $\Lambda_\mu - E_0 - R(E_s)$ as the root Λ_μ is close to E_s.

Further;

$$a_0 e^{-iE_s t} = \sum_{\mu=0}^{M} |A_0^\mu|^2 \,.\, e^{-i\Lambda_\mu t}$$

when the initial conditions are

$$\left. \begin{aligned} &a_0(0) = 1, \\ &a_r(0) = 0 \qquad (r = 1, 2, \ldots N). \end{aligned} \right\} \quad \ldots\ldots(13)$$

so

$$a_0 e^{-iE_s t} \simeq \int_{E_1}^{E_M} |A_0^\mu|^2 \,.\, \frac{1}{\Delta E} \,.\, e^{-iE_s t} dE_s, \quad \ldots\ldots(27)$$

and therefore (26) gives the form of the emitted line. This expression agrees with that of Weisskopf and Wigner when the term $R(E_s)$ is neglected. If this term were independent of E_s, equation (26) would give the usual form of line centred around E_0+R. However, for values of E close to the lower limit of the energy range, $R(E)$ is negative, while it is positive close to the upper limit of the energy range. For practical cases it is large and negative in the region of E_0. A rough estimate got by assuming

$$\begin{aligned}\int |H_{0r_1}|^2\, d\Omega &= \text{const}, \; \epsilon_r<\epsilon' \\ &= 0 \qquad \epsilon_r>\epsilon',\end{aligned}$$

where $\epsilon' = \boldsymbol{h}\boldsymbol{c}/a_0$ (a_0 = Bohr radius), gives a shift of the centre of the line of the order of $10^6 \,.\, \Gamma \boldsymbol{h}$ in the direction of decreasing energy, where $\Gamma\boldsymbol{h}$ is the line breadth. This shift of the line was first pointed out by Dirac (1927) and Oppenheimer (1930).

To give agreement with experiment the shift is, of course, neglected. It is, however, of some interest to note that it is connected with the conservation of energy. The usual expression for the intensity emitted in the frequency range, ν, $\nu+d\nu$ is

$$I(\nu)\, d\nu = \frac{\Gamma}{2\pi} \frac{\boldsymbol{h}\nu\, d\nu}{(\nu-\nu_0)^2+\Gamma^2/4},$$

where ν_0 is the resonant frequency. Clearly $\int_0 I(\nu)\, d\nu \to \infty$. Further, this expression gives too low a value for frequencies greater than ν_0 (due to neglect of the varying weight factor) and even if the energy of the emitted photons is cut off at ϵ' the energy discrepancy is considerable.*

The excited state of the atom and the states containing photons do not enter the fundamental equation (11) in a symmetrical fashion. The former interacts with all the latter, while each one of the latter states only interacts with one state—the excited state of the atom. This cause one root Λ_0 of the equation (14) to be a very large negative number, whereas the other roots are close to the unperturbed energies of the states containing photons. When the energy of the initial excited states is thus depressed, a shift in the position of the line is to be expected.

It is possible to extend the set (11) so that the equations become more "symmetrical". Consider briefly the interaction between the following states:—

ψ_0—the atom in the excited state, no photons;
ψ_{1i}—the atom in the normal state, with a photon of energy ϵ_i present;
ψ_{0ij}—the atom in the excited state with two photons of energy ϵ_i, ϵ_j present.

These states do not interact with the states ψ_1, ψ_{0i}, ψ_{1ij}, if the dipole interaction alone is considered. The fundamental equations then become (with an obvious notation)

$$\left.\begin{aligned}&(E_0-\Lambda)A_0+\sum_i H_{0,1i}A_{1i}=0,\\ &(E_{1j}-\Lambda)A_{1j}+H_{1j,0}A_0+\sum_i H_{1j,0ij}A_{0ij}=0,\\ &(E_{0ij}-\Lambda)A_{0ij}+H_{0ij,1i}A_{1j}+H_{0ij\,1}A_{1j}=0.\end{aligned}\right\} \quad \ldots\ldots(28)$$

---

* With the formula above

$$\int^{c/a_0} I(\nu)\, d\nu = \boldsymbol{h}\nu_0 + 1{\cdot}2\, \Gamma\boldsymbol{h}.$$

On eliminating the amplitudes A_{0ij}, and writing $H_{0ij,1j}=H_{0j,1}$, etc., equations (28) become

$$\left.\begin{aligned}\left\{E_{1j}-\Lambda-\sum_i\frac{|H_{1,0i}|^2}{E_{0ij}-\Lambda}\right\}A_{1j}-\sum_i\frac{H_{1j,0}H_{0\,1i}}{E_0-\Lambda}A_{1i}-\sum_i\frac{H_{1,0i}H_{0j,1}}{E_{0ij}-\Lambda}A_{1i}=0\\ (E_0-\Lambda)A_0+\sum_i H_{0,1i}A_{1i}=0.\end{aligned}\right\}\quad\ldots\ldots(29)$$

The original set (11) can be written in a similar form

$$\left.\begin{aligned}(E_{1j}-\Lambda)A_{1j}-\sum_i\frac{H_{1j,\,0}H_{0,1i}}{E_0-\Lambda}A_{1i}=0,\\ (E_0-\Lambda)A_0+\sum_i H_{0,1i}A_{1i}=0.\end{aligned}\right\}\quad\ldots\ldots(30)$$

If Λ is much less than the range of energy values E_{0ij} $(i=1,2,\ldots.M)$ the last term in the first equation of (29) can be neglected compared with the preceding term. Equations (30) lead to a solution in which the centre of the line, and therefore the most important values of Λ, are given by

$$\Lambda\sim E_0+R(E_0)=E_0+\int_{(E_0)}(E_0-E_{1i})^{-1}\,.\int|H_{0,1i}|^2\,d\Omega\,.\,\rho(E)_{1i}dE_{1i}$$

(neglecting the variation of $R(E)$ with E). Thus (29) will lead to a solution in which the centre of the line is given by

$$E_{1j}+\sum_i\frac{|H_{0\,1i}|^2}{E_{1j}-E_{0ij}}\simeq E_0+R(E_0).\qquad\ldots\ldots(31)$$

The difference between the most important E_{1j}, and E_0 is now reduced to a term of the form $\sum_i\frac{|H_{0,1i}|^2}{(E_0-E_{1i})^2}$. This suggests that the divergence difficulties in the emission case *may* be eliminated by taking account of higher-order processes which appear to have no physical importance. In equations (28) terms have been added which allow the atom to jump from the normal state to the excited state *emitting* a photon, and to jump back to the normal state *absorbing* a photon. Presumably the last term in the first of equations (29) modifies the shape of the line so as to reduce the energy discrepancy which arises when the shift is neglected.

§6. THE SOLUTION OF THE SCATTERING PROBLEM

The fundamental equations are

$$(E_s-\Lambda_t)A^t_s+\sum_{r}{}_{(s)}H_{sr}A^t_r=0\quad(s,t=0,1,2,\ldots M).\qquad\ldots\ldots(18)$$

Substituting $(\Lambda_t-E_s)A^t_s=V_{st}$ gives

$$V_{st}=\sum_r\frac{H_{sr}V_{rt}}{\Lambda_t-E_r}\qquad\ldots\ldots(32)$$

(where $H_{ss}=0$) provided $\Lambda_t\neq E_r$ for any t,r. It will be seen that all the $(M+1)$ roots Λ_t can be found in spite of this latter assumption. The r and s in (32) must include all variables required to describe the states (i.e. spins and polarizations as well as the momentum and energy).

The variation of H_{rs} with the energy of those states which are of interest will be small, and it can be neglected, if only those final states which lie in a small energy region are considered. Then, if l_1 and l_2 denote states with parallel directions * and different energies,

$$V_{l_1t}=\sum_r \frac{H_{l_1r}V_{rt}}{\Lambda_t-E_r} \qquad V_{l_2t}=\sum_r \frac{H_{l_2r}V_{rt}}{\Lambda_t-E_r}\ ;$$

so $V_{l_1t}=V_{l_2t}$, i.e. the function V_{ls} depends only on the "direction" of ψ_l and is independent of E_l. Thus (32) becomes

$$V_{lt}=\sum_{E_r} \frac{1}{\Lambda_t-E_r} \sum_{[r]} H_{lr}V_{rt}, \qquad \text{......(33)}$$

where $\sum_{[r]}$ denotes the sum over all states ψ_r belonging to any one energy level, and $\sum_{E_r}$ denotes the sum over the distinct energy levels. Equation (33) separates into an angular and an energy equation.

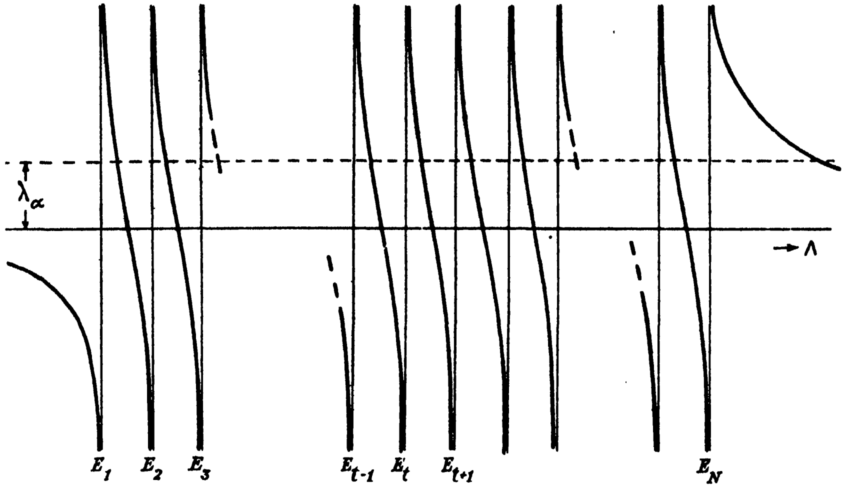


Figure 2. The intersections of the line λ_α and the curve $\sum_{E_l} \frac{1}{\Lambda-E_l}$.

Solutions of the equation

$$V_{l\alpha}=\lambda_\alpha \,.\, \sum_{r=1}^{m.\eta} H_{lr}V_{r\alpha}, \qquad \text{......(34)}$$

where m is the number of momentum values for each energy value, η is the spin and polarization degeneracy of each state, and λ_α is a real eigenvalue, are solutions of (33). The roots Λ_t are determined from (33) by

$$\lambda_\alpha=\sum_{E_r} \frac{1}{\Lambda_t-E_r}\,. \qquad \text{......(35)}$$

Figure 2 indicates the graphical solution of equation (35). Between each adjacent pair of energy levels E_r, E_{r+1}, there are as many roots Λ_t as there are finite eigenvalues λ_α. If λ_α is infinite, i.e. if solutions of the equation $\sum_{r=1}^{m.\eta} H_{lr}V_{r\alpha}=0$ exist; the corresponding Λ are identical with the energy levels E_r. A small λ_α ($\lambda_\alpha \simeq 1/\Delta E$ where $\Delta E=E_{r+1}-E_r$) leads to roots which lie roughly midway between the E_r.

---

* "Directions" must include spin and polarization variables.

If there are N distinct energy levels E_r, then $m \, . \, \eta \, . \, N = M+1$ (the total number of states) and as equation (34) in general has $m\eta$ eigenvalues λ_α, there are in general $M+1$ roots Λ_t. The last $m\eta$ roots of (35) are less than the lowest E_r or greater than the highest one.

It can be assumed that the m directions $[r]$ are uniformly distributed over the solid angle 4π, so that there are $(m/4\pi) \, . \, d\Omega_r$ states in the element of solid angle $d\Omega_r$. If m is very large, equation (34) can, with little error, be written in the form

$$V'_{l\zeta\alpha} = \lambda_\alpha \, . \, \frac{m}{4\pi} \sum_{\xi=1}^{\eta} \int H_{l\zeta r_\xi} V'_{r_\xi\alpha} d\Omega_r, \qquad \ldots\ldots(36)$$

where ψ_{r_ξ}, ψ_{r_ζ} are the various states with momentum r, and different spins or polarizations.

If $\eta \, . \, \rho(E) \, . \, dE \, . \, d\Omega_r$ is the number of states in the energy range E, $E+dE$ whose momenta lie in the solid angle $d\Omega_r$, then $\rho(E)\Delta E = m/4\pi$. Further, if $H'_{l_\zeta r_\xi}$ is the value of the matrix element for a box of volume unity, and $\rho'(E)$ is the corresponding density function, then $\rho(E) H_{l_\zeta r_\xi} = \rho'(E) \, H'_{l_\zeta r_\xi}$ and equation (36) becomes

$$V'_{l\zeta\alpha} = (\lambda_\alpha \, \Delta E) \rho'(E) \sum_{\xi=1}^{\eta} \int H'_{l\zeta r_\xi} V'_{r_\xi\alpha} \, d\Omega_r. \qquad \ldots\ldots(37)$$

Thus $(\lambda_\alpha \, . \, \Delta E)$ is independent of ΔE.

It can be readily seen that for the non-relativistic scattering of photons by an electron

$$\lambda_\alpha \, . \, \Delta E = O(1) \, . \, \frac{8\pi^2}{1+\mu^2/E^2} \, . \, (\hbar c/e^2),$$

where E is the total energy, and $O(1)$ denotes a constant of the order of magnitude of unity. Thus for weak coupling between the field and the scatterer $|\lambda \, \Delta E| \gg 1$. It is the contrary case of strong coupling when $|\lambda_\alpha \Delta E| = O(1)$, which shows important differences between the usual approximate solution and the exact solution.

It will be sufficient to show that the usual perturbation method (even allowing for degeneracy) fails when applied to the fundamental equations (18) unless $|\lambda_\alpha \Delta E| \gg 1$. A matrix $\boldsymbol{S}$ is required such that

$$\boldsymbol{S}^{-1} \, . \, \boldsymbol{K} \, . \, \boldsymbol{S} = \boldsymbol{W},$$

where $K_{rs} = E_r \delta_{rs} + H_{rs}$, and $\boldsymbol{W}$ is to be diagonal. Writing $H_{rs} = e^2 H^{11}_{rs}$ where e^2 is a small parameter, an approximate solution has to be found of the form $S = S_0 + e^2 S_1 + e^4 S_2 + \ldots$, $W = W_0 + e^2 W_1 + e^4 W_2 + \ldots$. The zero approximation is $H_0 S_0 = S_0 W_0$, where $(H_0)_{rs} = E_r \delta_{rs}$, and $W_0 = H_0$. This merely shows that S_0 reduces into matrices corresponding to the states in each distinct energy level. The first-order approximation is $H_0 S_1 - S_1 H_0 + H^{11} S_0 = S_0 W_1$. Denoting states belonging to the energy level E_a by the suffix a this gives

$$\sum_{p_a} H_{n_a p_a} \, . \, S_0(p_a m_a) = S_0(n_a m_a) \, . \, (W_1)_{m_a} \, . \, e^2 \qquad \ldots\ldots(38)$$

Equations (38) and (34) are identical so $S_0(p_a m_q) = V_{p, m_a}$ and $e^2 (W_1)_{m_a} = 1/\lambda_{m_a}$. Thus

$$e^2 (W_1)_{m_a} / \Delta E = 1/(\lambda_{m_a} \Delta E). \qquad \ldots\ldots(39)$$

It is clear from consideration of equation (35) or figure 2 that $e^2 (W_1)_{m_a}$ only gives a reasonable approximation to the actual energy perturbation when $|\lambda_{m_a} \, . \, \Delta E| \gg 1$. Otherwise (39) gives a quite incorrect result.

§7. DEGENERACY AND NORMALIZATION

The basic integral equation (37) may be degenerate. For example, if $H'_{lr}=H$, a constant, then there is only one solution $V_{r_1}=\text{const.}$, $\lambda_1 \Delta E = 1/H \,.\, \rho'(E) \,.\, 4\pi$; so there is an infinity of linearly independent degenerate solutions, with infinite eigenvalues, satisfying $\int V_{r\alpha} d\Omega_r = 0$. In this case an infinity of roots satisfy $\Lambda = E_t$. However, this type of degeneracy is trivial and can be easily removed.

Normalize the solutions of (37) so that

$$\sum_{\xi=1}^{\eta} \int \overline{V}'_{l_\xi\alpha} V_{r_\xi\beta} d\Omega_r = \delta_{\alpha\beta}. \qquad \ldots\ldots(40)$$

Then by a well known theorem H'_{lr} can be expressed in the form

$$\rho'(E) H'_{l_\zeta r_\xi} = \sum_{\alpha=1}^{\infty} V'_{l_\zeta\alpha} \overline{V}'_{r_\xi\alpha}/(\lambda_\alpha \Delta E), \qquad \ldots\ldots(41)$$

provided the series is uniformly convergent. Conversely if the $V'_{l_\zeta\alpha}$ are independent normal orthogonal functions, in the sense of (40), the solutions of the integral equations (37) in which the kernel is given by (41), are the $V'_{l_\zeta\alpha}$, and the corresponding eigenvalues are the $\lambda_\alpha \,.\, \Delta E$.

Assume for simplicity that in the degenerate case above there is no spin or polarization degeneracy. Replace the interaction H'_{lr} by

$$\rho'(E) H'_{lr} = \rho'(E) H + \sum_{i=1}^{\infty} Y_i(l) \overline{Y}_i(r)/b_i,$$

where $Y_i(l)$ is the ith spherical harmonic, and the b_i are a series of constants which increase in magnitude very rapidly as i increases, and b_1 is as large as desired, Then the difference between H'_{lr} and H can be made extremely small, and cannot affect the physical problem appreciably. The eigenvalues $\lambda_\alpha \Delta E$ of the integral equation will then all (except the first) be very large and so will correspond to extremely weak coupling. (It will also be seen from the final stages that the influence of the b_i is negligible.)

It is also apparent, by similar reasoning, that the transition from the linear equations (34) to the integral equation (37) is valid, provided that the change in H_{lr}, in going from any direction to any "adjacent" direction, is small compared with the value of H_{lr}.

(a) *Orthogonal properties and normalization*

The first condition (6) requires

$$\sum_{r=0}^{M} \overline{A}_r^s A_r^{s'} = \delta_{ss'};$$

or

$$\sum_{r=0}^{M} \frac{\overline{V}_{rs} V_{rs'}}{(\Lambda_s - E_r)(\Lambda_{s'} - E_r)} = \delta_{ss'}. \qquad \ldots\ldots(42)$$

Putting $s=t,\alpha$; $s'=t',\alpha'$; (41) can be written

$$\sum_{E_r} \frac{1}{(\Lambda_{t\alpha} - E_r)(\Lambda_{t'\alpha'} - E_r)} \cdot \sum_{r=1}^{m\eta} \overline{V}_{r,t\alpha} V_{r,t'\alpha'} = \delta_{tt'} \,.\, \delta_{\alpha\alpha'}, \qquad \ldots\ldots(42\,a)$$

where $\Lambda_{t\alpha}$ $(t=1, 2, \ldots N)$ denote the roots of

$$\lambda_\alpha = \sum_{E_r} \frac{1}{\Lambda - E_r}.$$

It can easily be seen that (42 *a*) is satisfied, as follows:—

(i) from (35) it follows that

$$\sum_{E_r} \frac{1}{(\Lambda_{t\alpha}-E_r)(\Lambda_{t'\alpha}-E_r)}=0,$$

if $E_t \neq E$, i.e. if $t \neq t'$;

(ii) equation (34) has orthogonal solutions, on

$$\sum_{r=1}^{m\eta} \overline{V}_{r\alpha} V_{r\alpha'}=0, \quad \text{if } \alpha \neq \alpha'.$$

The V functions are normalized by (42), so that

$$\sum_{E_r} \frac{1}{(\Lambda_{t\alpha}-E_r)^2} \sum_{r=1}^{m\eta} |V_{r,t\alpha}|^2=1. \qquad \ldots\ldots(43)$$

It is more useful to express (43) in terms of the solutions V' of the integral equation. To be more precise let $Z_{l_\xi\alpha}$ be the *normal* orthogonal solutions of (34) and $V'_{l_\xi\alpha}$ the normal orthogonal solutions of (37).

Then

$$\left.\begin{aligned} &\sum_{\xi=1}^{\eta} \sum_{l=1}^{m} \overline{Z}_{l_\xi\alpha} Z_{l_\xi\alpha'} = \delta_{\alpha\alpha'} \\ &\sum_{\alpha=1}^{m\,.\,\eta} \overline{Z}_{l_\zeta\alpha} Z_{l'_\xi\alpha} = \delta_{ll'} \cdot \delta_{\alpha\alpha'} \end{aligned}\right\} ; \qquad \ldots\ldots(44)$$

and

$$\sum_{\xi=1}^{\eta} \int \overline{V}'_{l_\xi\alpha} V'_{l_\xi\alpha'}\, d\Omega_l = \delta_{\alpha\alpha'}. \qquad \ldots\ldots(45)$$

Taking $Z_{l_\xi\alpha}=a_\alpha V_{l_\xi\alpha}$, where a_α only depends on α, the first equation of (44) gives

$$|a_\alpha|^2 \sum_{\xi=1}^{\eta} \sum_{l=1}^{m} |V'_{l_\xi\alpha}|^2 = (m/4\pi) \sum_{\xi} \int |V'_{l_\xi\alpha}|^2\, d\Omega_l = 1$$

so $|a_\alpha|^2 = 4\pi/m$.

(The second equation of (44) then takes the form

$$(4\pi/m) \,.\, \sum_{\alpha=1}^{m\eta} \overline{V}'_{l_\zeta\alpha} V'_{l'_\xi\alpha} = \delta_{ll'} \,.\, \delta_{\zeta\xi},$$

and as $m \to \infty$ this becomes

$$\lim_{m\to\infty} (4\pi/m) \,.\, \sum_{\alpha=1}^{m\eta} \overline{V}'_{l_\zeta\alpha} V'_{l'_\xi\alpha} = \delta_{ll'} \,.\, \delta_{\zeta\xi},$$

the form of the second orthogonality condition for continuous eigen-functions.)

If the $V_{l,t\alpha}$ which satisfies (43) is related to $Z_{l,\alpha}$ by

$$|V_{l,t\alpha}|^2 = |b_{t\alpha}|^2 \,.\, |Z_{l,\alpha}|^2,$$

then (43) gives

$$|b_{t\alpha}|^2 \sum_{E_r} \frac{1}{(\Lambda_{t\alpha}-E_r)^2} = 1, \qquad \ldots\ldots(46)$$

determining the $|b_{t\alpha}|^2$.

Finally,

$$A^s_{r\zeta} = \frac{1}{(\Lambda_{t\alpha}-E_r)} \cdot b_{t\alpha} \cdot \sqrt{4\pi/m} \,.\, V'_{r_\zeta\alpha}, \qquad \ldots\ldots(47)$$

where $s = t, \alpha$.

Before applying the series for $\cot z$ and $\operatorname{cosec}^2 z$ to the solution it is worth remembering that the number of final energy levels may be very large, but the energy range which these levels cover has to be much less than the difference between the initial energy and the nearest of the intermediate energies. The energy range of the final states cannot be very great or the method of solution used at the beginning of this section (the assumption that H_{rs} is independent of the energy) will not in general be valid. As the "line breadth" must decrease as $1/L^3$, this restriction cannot affect the physically true solution, but it may mean that some divergences which would otherwise arise are not apparent. However, allowing the breadth of the band of final energy states to increase should show up some at least of the divergences.

Equation (35) can be solved in a similar way to (14), viz.:—

$$\lambda_\alpha = \sum_{r=-n}^{+n} \frac{1}{(\Lambda_{t\alpha}-E_s)-r\Delta E} + \left(\sum_{r=1}^{s-n-1} + \sum_{r=s+n+1}^{N}\right)\frac{1}{\Lambda_{t\alpha}-E_r},$$

giving the value of $\Lambda_{t\alpha}$ in the vicinity of E_s. So

$$\lambda_\alpha = \frac{\pi}{\Delta E} \cdot \cot\{(\Lambda_{t\alpha}-E_s)\pi/\Delta E\} + \frac{1}{\Delta E}\int_{(E_s)} \frac{1}{(E_s-E_r)}\,dE_r, \qquad \ldots\ldots(48)$$

using the notation of § 5. Thus

$$\lambda_\alpha \,.\, \Delta E = \pi \cot\{(\Lambda_{t\alpha}-E_s)\pi/\Delta E\} + \sigma(E_s),$$

where

$$\sigma(E_s) = \int_{(E_s)} \frac{1}{E_s-E_r}\,dE_r\,.$$

$\sigma(E_s)$ is clearly similar to the term $R(E_s)$ of § 5. If there were no intermediate states and H_{rs} could be taken as independent of energy this term σ could become a very large negative quantity.

Equation (46) for $|b_{t\alpha}|^2$ is treated similarly to equation (21). In the vicinity of E_s

$$\sum_{E_r} \frac{1}{(\Lambda_{t\alpha}-E_r)^2} \simeq \sum_{r=-\infty}^{+\infty} \frac{1}{(\Lambda_{t\alpha}-E_s-r\Delta E)^2} = (\pi/\Delta E)^2 \operatorname{cosec}^2\{(\Lambda_{t\alpha}-E_s)\pi/\Delta E\}.$$

Here the remainder terms for large r can be neglected in a similar fashion to those arising from equation (21). Using (48) gives

$$\sum_{E_r} \frac{1}{(\Lambda_{t\alpha}-E_r)^2} = \frac{1}{(\Delta E)^2}\{\pi^2 + (\lambda_\alpha \Delta E - \sigma(E_s))^2\}$$

so

$$|b_{t\alpha}|^2 = \frac{(\Delta E)^2}{\pi^2 + (\lambda_\alpha \Delta E - \sigma(E_s))^2} \qquad \ldots\ldots(49)$$

$|b_{t\alpha}|^2$ gives the measure of the importance of the eigenvalue $(\lambda_\alpha \Delta E)$ in the solution $A_r^{t\alpha}$. $\sigma(E_s)$ must be neglected in (49) else $|\lambda_\alpha \Delta E| = O(1)$ may lead to small contributions, while $\lambda_\alpha \Delta E \gg 1$ could give a large $|b_{t\alpha}|^2$. Thus there is a divergent term* arising from the H_{rs}—the second order matrix elements in equation (18). Putting $\sigma(E_s) = 0$ gives†

$$|b_{t\alpha}|^2 = \frac{(\Delta E)^2}{\pi^2 + (\lambda_\alpha \Delta E)^2}\,. \qquad \ldots\ldots(50)$$

---

* This divergence difficulty cannot be removed by the present methods.

† Equations (47) and (50) show that the "line breadth" behaves as $1/(A_t - E)^2$, so there is no breadth in the sense of the emission case.

(b) *The variation of the initial state*

If the initial state is ψ_{k_ξ} the initial conditions are

$$\left.\begin{aligned} a_{k_\xi}(0) &= 1, \\ a_s(0) &= 0 \qquad s \neq k_\xi, \end{aligned}\right\} \qquad \ldots\ldots(51)$$

so $c_\mu = \bar{A}^{\mu}_{k_\xi}$, and for any state ψ_{r_ξ} the time variation is given by

$$\begin{aligned} a_{r_\zeta} e^{-iE_r t} &= \sum_{\mu=0}^{M} \bar{A}^{\mu}_{k_\xi} A^{\mu}_{r_\zeta} e^{-i\Lambda_\mu t} \\ &= \sum_{\alpha=1}^{m.\eta} \sum_{p=1}^{N} \bar{A}^{p,\alpha}_{k_\xi} A^{p,\alpha}_{r} e^{-i\Lambda_{p\alpha}.t} \qquad (\mu = p, \alpha). \end{aligned}$$

n particular

$$a_{k_\xi} e^{-iE_k t} = \sum_{\alpha=1}^{m.\eta} \sum_{p=1}^{N} | A^{p,\alpha}_{k_\xi} |^2 e^{-i\Lambda_{p\alpha}.t},$$

o

$$a_{k_\zeta} = \sum_{\alpha} \sum_{p} | A^{p,\alpha}_{k_\xi} |^2 \,.\, e^{-i(n\Delta E + \delta_\alpha)t},$$

vhere $\Lambda_{p\alpha} - E_k = n\Delta E + \delta_\alpha$. Equation (49) states

$$(\pi/\Delta E) \cot(\pi\delta_\alpha/\Delta E) = \lambda_\alpha,$$

hus, using (47)

$$a_{k_\xi} \simeq \frac{4\pi}{m} \sum_{\alpha} \frac{| V'_{k_\xi\alpha} |^2}{\pi^2 \operatorname{cosec}^2(\pi\delta_\alpha/\Delta E)} \,.\, (\Delta E)^2 \sum_{n=-\infty}^{+\infty} \frac{1}{(n\Delta E + \delta_\alpha)} e^{-i(n\Delta E + \delta_\alpha)t} \ldots\ldots\ldots(52)$$

ubstituting $t \,.\, \Delta E = x$, the last summation becomes

$$\frac{1}{(\Delta E)^2} \sum_{n=-\infty}^{+\infty} \frac{1}{(n + \delta_\alpha/\Delta E)^2} \,.\, e^{-i(n + \delta_\alpha/\Delta E).x}.$$

his sum is periodic in x with period 2π. In Appendix I it is evaluated. The sum a continuous function of x, but the first derivative of the sum is discontinuous at $= s\pi$, $s = 0, \pm 1, \pm 2, \ldots$ For $0 \leqslant x \leqslant \pi$ (52) gives

$$a_{k_\xi} = \frac{4\pi}{m} \sum_{\alpha=1}^{m.\eta} | V'_{k_\xi\alpha} |^2 \left\{ 1 - \pi x \frac{1 - i\cot(\pi\delta_\alpha/\Delta E)}{\pi^2 \operatorname{cosec}^2(\pi\delta_\alpha/\Delta E)} \right\}. \qquad \ldots\ldots(53)$$

Vhen $| a_k |^2$ is evaluated the imaginary terms in (53) will give rise to terms in x^2 only. f $x \ll 2\pi$, i.e. if $t \ll 2\pi/\Delta E$, the x^2 terms can be neglected. As $| a_{k_\xi} |^2$ is only to be valuated correctly for the constant terms and those depending on x, it is sufficient take

$$a_{k_\xi} = \frac{4\pi}{m} \sum_{\alpha=1}^{m.\eta} | V'_{k_\xi\alpha} |^2 \left\{ 1 - \frac{\pi x}{\pi^2 \operatorname{cosec}^2(\pi\delta_\alpha/\Delta E)} \right\}. \qquad \ldots\ldots(54)$$

Thus the transition probability Γ is given by

$$\Gamma/2 = \Delta E \,.\, \frac{4\pi}{m} \,.\, \sum_{\alpha} | V_{k_\xi\alpha} |^2 \frac{\pi}{\pi^2 \operatorname{cosec}^2(\pi\delta_\alpha/\Delta E)},$$

or

$$\rho(E) \,.\, \Gamma/2 = \sum_{\alpha} | V'_{k_\xi\alpha} |^2 \,.\, \frac{\pi}{\pi^2 + (\lambda_\alpha \Delta E)^2}. \qquad \ldots\ldots(55)$$

Remembering that the $V'_{k\alpha}$ are normalized functions it is obvious that in general the higher eigenvalues $\lambda_\alpha \Delta E$ give very little contribution to the transition probability, so in practice only a few of the lowest solutions of (37) need be considered.

(c) *The scattered states*

The analogue of (52) for any state a_r $(r \neq k)$ is

$$a_{r_\zeta} = \frac{4\pi}{m} \sum_\alpha \overline{V}'_{k_\xi\alpha} V'_{r_\zeta\alpha} \frac{1}{\pi^2 \operatorname{cosec}^2(\pi\delta_\alpha/\Delta E)}$$

$$\times \sum_{n=-\infty}^{+\infty} \frac{1}{(n+q+\delta_\alpha/\Delta E)(n+\delta_\alpha/\Delta E)} \cdot e^{-i(n+q+\delta_\alpha/\Delta E).x}$$

where $\Lambda_{p\alpha} - E_r = (\Lambda_{p\alpha} - E_k) + (E_k - E_r) = (n\Delta E + \delta_\alpha) + q\Delta E$, $E_k - E_r = q\Delta E$. In the Appendix it is shown that if $q \neq 0$

$$\sum_{n=-\infty}^{+\infty} \frac{1}{(n+q+\delta_\alpha/\Delta E)(n+\delta_\alpha/\Delta E)} \cdot e^{-i(n+q+\delta_\alpha/\Delta E).x} = \frac{\pi}{q}(e^{-iqx}-1)\{1+\cot(\pi\delta_\alpha/\Delta E)\}$$

so if $q \neq 0$ (i.e. $E_r \neq E_k$),

$$a_{r_\zeta} = \frac{4\pi}{m} \sum_\alpha \overline{V}'_{k_\xi\alpha} V'_{r_\zeta\alpha} \cdot \frac{1}{\pi\{-i+\cot(\pi\delta_\alpha/\Delta E)\}} \cdot \frac{1}{q}(e^{-iqx}-1);$$

and if $E_r = E_k$,

$$a_{r_\zeta} = \frac{4\pi}{m} \sum_\alpha \overline{V}'_{k_\xi\alpha} V'_{r_\zeta\alpha} \cdot \frac{i}{\pi\{-i+\cot(\pi\delta_\alpha/\Delta E)\}} \cdot x.$$

as

$$\frac{4\pi}{m} \sum_\alpha \overline{V}'_{k_\xi\alpha} V'_{r_\zeta\alpha} = 0, \quad \text{if } r \neq k.$$

Only the total probability of scattering in a given "direction" is of interest, so the probabilities for all the states of different energies having this "direction" must be added. This gives

$$\sum_{E_r} |a_r|^2 = \left| \frac{4\pi}{m} \sum_\alpha \overline{V}'_{k_\xi\alpha} V'_{r_\zeta\alpha} \frac{i}{\pi\{-i+\cot(\pi\delta_\alpha/\Delta E)\}} \right|^2 \cdot \left\{ x^2 + \sum_q{}' \frac{1}{q^2} |e^{-iqx}-1|^2 \right\},$$

where $\sum_q{}'$ denotes summation over all integers except zero. Further,

$$x^2 + \sum_q{}' \frac{1}{q^2} |e^{-iqx}-1|^2 = x^2 + 2\sum_q{}' \frac{1-\cos qx}{q^2} = 2\pi x$$

(see Appendix). Thus

$$\sum_{E_r} |a_{r_\zeta}|^2 = \left| \frac{4\pi}{m} \sum_\alpha \overline{V}'_{k_\xi\alpha} V'_{r_\zeta\alpha} \frac{1}{\pi + i(\lambda_\alpha \Delta E)} \right|^2 . \, 2\pi \, . \, \Delta E \, . \, t. \quad \ldots\ldots(56)$$

If $\sigma(E_s)$ were not neglected in (50) the transition probability would become

$$\rho(E) \, . \, \Gamma/2 = \sum_\alpha |V'_{k_\xi\alpha}|^2 \frac{\pi}{\pi^2 + \{\lambda_\alpha \Delta E - \sigma(E)\}^2};$$

so, except in the region of initial energy E where $\sigma(E) \simeq 0$, the physically unimportant eigenvalues satisfying $|\lambda_\alpha \Delta E| \gg 1$ might give much greater contributions than eigenvalues satisfying $|\lambda_\alpha \Delta E| = O(1)$.

In the evaluation of the series $\sum_{E_r} \frac{1}{(\Lambda_{p\alpha} - E_r)^2}$, which gave $|b_{p\alpha}|^2$, it is necessary (as in the emission case) to assume that the summation is cut off at some arbitrarily large energy value,

§8. THE HEITLER-PENG INTEGRAL EQUATION

The Heitler-Peng integral equation

$$U_{r_\zeta k_\xi} = H'_{r_\zeta k_\xi} + i\pi\rho'(E) \sum_{\xi'=1}^{\eta} \int H'_{r_\zeta t_{\xi'}} U_{t_{\xi'} k_\xi} \, d\Omega_t, \qquad \ldots\ldots(57)$$

now appears as a useful method for making the summations occurring in the expressions (55) and (56).

From the homogeneous linear integral equation (37), whose solutions satisfy the normalizing conditions (40), it is possible to construct a non-homogeneous linear integral equation

$$U_{r_\zeta} = F_{r_\zeta} + g\rho'(E) \sum_{\xi'=1}^{\eta} \int H'_{r_\zeta t_{\xi'}} U_{t_{\xi'}} \, d\Omega_t, \qquad \ldots\ldots(58)$$

where g is, at present, an unknown constant, and F_{r_ζ} an unkown function of r_ζ.

The solution U_{r_ζ} of (58) can be made to give the $\sum\limits_\alpha$ term in (56).

If $U_{r_\zeta} = \sum\limits_\alpha u_\alpha V'_{r_\zeta\alpha}$ $F_{r_\zeta} = \sum\limits_\alpha f_\alpha V'_{r_\zeta\alpha}$, the solution of (58) is given by

$$u_\alpha = \mu_\alpha f_\alpha / (\mu_\alpha - g), \quad \text{where } \mu_\alpha = \lambda_\alpha \Delta E.$$

Putting $\rho'(E)\mu_\alpha f_\alpha = \overline{V}'_{k_\xi\alpha}$, $g = i\pi$ gives

$$\rho'(E) U_{r_\zeta} = \rho'(E) \sum_\alpha u_\alpha V'_{r_\zeta\alpha} = i \sum_\alpha \frac{\overline{V}'_{k_\xi\alpha} V'_{r_\zeta\alpha}}{\pi + i(\lambda_\alpha \Delta E)}. \qquad \ldots\ldots(59)$$

Writing $U_{r_\zeta} = U_{r_\zeta k_\xi}$, the equation (56) becomes

$$\sum_{Er} |a_r|^2 = 2\pi x \,.\, \left| \frac{4\pi}{m} \rho'(E) U_{r_\zeta k_\xi} \right|^2.$$

The probability for scattering in the element of solid angle $d\Omega_r$ about r is given by

$$\rho(E) \,.\, \Delta E \,.\, d\Omega_r \sum_{Er} |a_{r_\zeta}|^2 = 2\pi t \,.\, \rho(E) \,.\, (\Delta E)^2 \left| \frac{4\pi}{m} \rho'(E) U_{r_\zeta k_\xi} \right|^2 . \, d\Omega_r.$$

Using the relation* $(4\pi/m) \,.\, \rho'(E) = (L^3 \,.\, \Delta E)^{-1}$, the transition probability γ for scattering into the solid angle $d\Omega_r$ becomes

$$\gamma = \frac{1}{L^3} \,.\, 2\pi\rho'(E) \left| U_{r_\zeta k_\xi} \right|^2 . \, d\Omega_r \quad (\hbar = 1), \qquad \ldots\ldots(60)$$

where $U_{r_\zeta k_\xi}$ satisfies equation (57).

The Heitler-Peng result for the transition probability of the initial state

$$\Gamma = \frac{1}{L^3} \,.\, 2\pi \mathscr{R} \left[\rho'(E) \sum_{\zeta=1}^{\eta} \int H'_{k_\xi r_\zeta} U_{r_\zeta k_\xi} \, d\Omega_r \right] \qquad \ldots\ldots(61)$$

can easily be verified.

$$\rho'(E) \sum_{\zeta=1}^{\eta} \int H'_{k_\xi r_\zeta} U_{r_\zeta k_\xi} \, d\Omega_r = \frac{1}{\rho'(E)} \,.\, \sum_{\zeta=1}^{\eta} \int \sum_\alpha \sum_\beta \frac{V'_{k_\xi\alpha} \overline{V}_{r_\zeta\beta}}{\mu_\alpha} \,.\, \frac{V'_{r_\zeta\beta} \overline{V}_{k_\xi\beta}}{\mu_\beta - i\pi} \,.\, d\Omega_r$$

using (59) and the expransion of H'_{kr} in terms of eigensolutions. Thus

$$\rho'(E) \sum_{\zeta=1}^{\eta} \int H'_{\xi r_\zeta} U_{r_\zeta k_\xi} \, d\Omega_r = \frac{1}{\rho'(E)} \left\{ \sum_\alpha \frac{|V'_{k_\xi\alpha}|^2}{\mu_\alpha + \pi^2} + i\pi \sum_\alpha \frac{|V'_{k_\xi\alpha}|^2}{\mu_\alpha(\mu_\alpha^2 + \pi^2)} \right\},$$

---

* As above, $\rho(E)$ and H_{rt} are the density function and matrix elements for a cubic box with edges of length L, while $\rho'(E)$ and H'_{tt} are the same quantities for a cubic box with unit edges.

so (61) and (55) are identical. It can, however, be seen that a simpler expression would be

$$\Gamma=\frac{1}{L^3}\cdot 2\pi\rho'(E)\sum_{\zeta=1}^{\eta}\int|U_{r_\zeta k_\xi}|^2\,d\Omega_r, \qquad \ldots\ldots(62)$$

which is also identical with (55). Those transitions which lead from the initial state ψ_{k_ξ} to other states with the same direction but different energies can be entirely neglected.

Gormley and Heitler (1944) have shown that the equations (60) and (61) have the correct Lorentz transformation properties. Thus (55) and (56) will give the transition probabilities when the total momentum of the system is not zero, and in cases in which the energy levels are no longer almost equi-distant, and almost equally degenerate.* In a new Lorentz frame of reference, H_{rs}, the interaction matrix, will have a different form, and consequently the λ_α and $V'_{r\alpha}$ will change in going from one frame of reference to another. It may be of advantage to choose the frame of reference in which H_{rs} is least dependent on the variables, as equation (37) can possibly be solved by approximate methods in that case.

Finally, it is worth noticing that equation (41) assumes that the correct normalization condition is $\sum_{r=0}^{M}\bar{A}_r^s A_r^{s'}=\delta_{ss'}$; while the correct condition is that the total probability of the system being in any of the final *or intermediate* states is unity. The treatment given in §§ 6 and 7 is thus only valid provided $\sum_i|A_i|^2 \ll 1$, where A_i is the amplitude of an intermediate state.

Using the second of equations (15) and the values of A_r given in § 7 it can be shown that this condition holds provided the initial energy lies close to the centre of the energy range of the final states. Otherwise it is not true; and the intermediate states thus give rise to a further divergence difficulty in the solution.

ACKNOWLEDGMENTS

I am indebted to Professor Dirac, and to Drs. B. Ferretti and H. W. Peng, for useful discussion.

APPENDIX

Various summations

1. The values of

$$\left.\begin{aligned} f(x)&=\sum_{n=-\infty}^{+\infty}\frac{e^{inx}}{n+a}\\ g(x)&=\sum_{n=-\infty}^{+\infty}\frac{e^{inx}}{(n+a)^2}\end{aligned}\right\}\quad(a\neq\text{an integer})$$

are required.

Now

$$\sum_{r=-n}^{+n}\frac{e^{irx}}{r+a}\cdot e^{iax}=i\int_0^x e^{iat}\sum_{r=-n}^{+n}e^{irt}.\,dt+\sum_{r=-n}^{+n}\frac{1}{r+a},$$

and

$$\sum_{r=-n}^{+n}e^{irt}=\frac{\sin(n+\frac{1}{2})t}{\sin(t/2)}.$$

---

* If $\rho'(E)$ depends on the angles it must be included in the kernels of the integral equations.

Further, if $0\leqslant x\leqslant\pi$,

$$\int_0^x \frac{\sin(n+\frac12)t}{\sin(t/2)}\, e^{iat}\,dt = 2\int_0^x e^{iat}\frac{\sin(n+\frac12)t}{t}\,dt + \int_0^x e^{iat}\sin(n+\tfrac12)t\left\{\frac{1}{\sin(t/2)} - \frac{1}{t/2}\right\}dt.$$

The last term equals

$$\left[e^{iat}\left(\frac{1}{\sin(t/2)}-\frac{1}{t/2}\right)\frac{\cos(n+\frac12)t}{n+\frac12}\right]_0^x$$

$$+\frac{1}{n+\frac12}\int_0^x \cos(n+\tfrac12)t\,.\,\frac{d}{dt}\left\{e^{iat}\left(\frac{1}{\sin(t/2)}-\frac{1}{t/2}\right)\right\}dt$$

and as $\dfrac{1}{\sin y}-\dfrac{1}{y}$, and $\dfrac{d}{dy}\left(\dfrac{1}{\sin y}-\dfrac{1}{y}\right)$ are bounded in the range $0\leqslant y\leqslant\pi/2$, these terms are $O(1/n)$. Thus

$$\sum_{r=-\infty}^{+\infty}\frac{e^{irx}}{r+a}\,.\,e^{iax}=\pi\{i+\cot a\pi\}$$

as

$$\sum_{r=-\infty}^{+\infty}\frac{1}{r+a}=\pi\cot a\pi.$$

Thus

$$f(x)e^{iax}=\pi\{i+\cot a\pi\} \qquad (0\leqslant x\leqslant\pi)$$

and

$$=\pi\{-i+\cot a\pi\} \qquad (-\pi\leqslant x\leqslant 0).$$

The function $f(x)$ obviously has period 2π.

Similarly, if $0\leqslant x\leqslant\pi$,

$$\sum_{r=-n}^{+n}\frac{e^{irx}}{(r+a)^2}\,.\,e^{iax}=i\int_0^x\sum_{r=-n}^{+n}\frac{e^{irt}}{r+a}\,.\,e^{iat}\,.\,dt+\sum_{r=-n}^{+n}\frac{1}{(r+a)^2}.$$

The series for $f(x)e^{iax}$ is uniformly convergent over $(0,\pi)$ except in the neighbourhood of 0 and π, and it is boundedly convergent over the whole interval, so we may write

$$\sum_{r=-\infty}^{+\infty}\frac{e^{irx}}{(r+a)^2}\,.\,e^{iax}=i\int_0^x\pi(i+\cot a\pi)dt+\sum_{r=-\infty}^{+\infty}\frac{1}{(r+a)^2} \qquad (0\leqslant x\leqslant\pi).$$

Thus

$$g(x)e^{iax}=i\pi(i+\cot a\pi)x+\pi^2\operatorname{cosec}^2 a\pi \qquad (0\leqslant x\leqslant\pi)$$

$$=i\pi(-i+\cot a\pi)x+\pi^2\operatorname{cosec}^2 a\pi \qquad (-\pi\leqslant x\leqslant 0).$$

2. The value of

$$\sum_{n=-\infty}^{+\infty}{}'\,\frac{e^{inx}}{n^2} \qquad (0\leqslant x\leqslant\pi).$$

As before,

$$\sum_{r=-\infty}^{+\infty}\frac{e^{irx}}{r}=\lim_{n\to\infty} i\int_0^x\sum_{r=-n}^{+n}{}'\,e^{irt}\,.\,dt+\sum_{r=-\infty}^{+\infty}{}'\,\frac{1}{r}$$

$$=\lim_{n\to\infty} i\int_0^x\left\{\frac{\sin(n+\frac12)t}{\sin(t/2)}-1\right\}dt$$

$$=i(\pi-x) \qquad (0\leqslant x\leqslant\pi).$$

Thus

$$\sum_{n=-\infty}^{+\infty}{}'\,\frac{e^{inx}}{n^2}=i\int_0^x i(\pi-x)dx+\sum_{n=-\infty}^{+\infty}{}'\,\frac{1}{n^2}$$

$$=-\pi x+\frac{x^2}{2}+\frac{\pi^2}{3}.$$

Further

$$\sum_{r=-\infty}^{+x}{}' \frac{1-\cos rx}{r^2} = \sum_r{}' \frac{1}{r^2} - \sum_r{}' \frac{e^{irx}}{r^2}$$

$$= \pi x - \frac{x^2}{2} \qquad (0 \leqslant x \leqslant \pi).$$

3. **The sum**

$$\sum_{n=-\infty}^{+\infty} \frac{e^{inx}}{(n+a)(n+b)} \qquad \text{if } a = m+b,$$

where m is a non-zero integer.

If $0 \leqslant x \leqslant \pi$, we have

$$\sum_{r=-n}^{+n} \frac{e^{irx} \,.\, e^{ibx}}{(r+a)(r+b)} = i\int_0^x \sum_{r=-n}^{+n} \frac{e^{irx}}{r+a} . e^{ibx} dx + \sum_{r=-n}^{+n} \frac{1}{(r+a)(r+b)} \,.$$

The latter sum is zero, and as before we have

$$\sum_{n=-\infty}^{+\infty} \frac{e^{inx}}{(n+a)(n+b)} \,.\, e^{ibx} = i\int_0^x \pi(i+\cot a\pi) e^{i(b-a)t} dt\,;$$

so

$$\sum_{n=-\infty}^{+\infty} \frac{e^{inx}}{(n+a)(n+b)} = \frac{\pi}{b-a}(i+\cot a\pi)(e^{-iax} - e^{-ibx}).$$

REFERENCES

Dirac, 1927. *Z.S.*, **44**, 585.
Gormley and Heitler, 1944. *Proc. Roy. Irish Acad.*, 50 A, **4**, 29.
Heitler, 1941. *Proc. Camb. Phil. Soc.*, **37**, 291.
Heitler and Peng, 1942. *Proc. Camb. Phil. Soc.*, **38**, 296.
Oppenheimer, 1930. *Phys. Rev.*, **35**, 461.
Peng, 1944. *Nature, Lond.*, **154**, 544.
Peng, 1946. *Proc. Roy. Soc.*, A, **186**, 119.
Weisskopf and Wigner, 1930. *Z.S.*, **63**, 54 ; 1930. *Ibid.*, **65**, 18.
Wilson, 1941. *Proc. Camb. Phil. Soc.*, **37**, 301.

A WAVEFRONT SHEARING INTERFEROMETER

By W. J. BATES,

H. H. Wills Physical Laboratory, University of Bristol

MS. received 26 *December* 1946

ABSTRACT. A new type of interferometer is described by means of which the asphericity of an optical wavefront can be measured, by testing it against itself with lateral displacement or shear. Continuous control of the amount of this shear, and of the meridional or sagittal fringes is obtained in white light.

§ 1. INTRODUCTION

THE possibility of determining uniquely the shape of a wavefront by interferometric examination on superposition with another wavefront seems not to have been exploited fully. The usual methods involve the provision of a substantially error-free wavefront as reference standard. When, however, one is allowed to displace one wavefront laterally with respect to the other, or to introduce a "shear" between them, then the error-free standard may be unnecessary; for the asphericities of two *dissimilar* wavefronts of revolution symmetry (about known centres), may be determined uniquely by a single interference pattern.

The existence of such solutions may be seen in the following way: Let W_1 and W_2 (figure 1) be two such wavefronts, giving an interferogram in their overlap region; and let the distance between the centres of revolution symmetry, C_1 and C_2, be greater than zero, and less than or equal to the distance AB. Then the asphericity of W_1 inside the circle α may be determined by the fringe intersections with the circle β around the arc EC_1F, this asphericity being referred to a sphere passing through β and C_2. The asphericity of W_2 inside the circle β may be obtained in a similar way, and the tilt and difference of curvature between these reference caps is also obtained. The process may be extended step by step along the line AB to include the whole of both wavefronts, with no assumptions as to absence of discontinuities. If the circles α and β intersect or touch the peripheries of W_1 and W_2, then the complete solution is obtained in a single step; and if C_1 and C_2 lie outside the interferogram then a solution is still possible in the interferogram region. It is clear also that when the distance C_1C_2 is zero then no solution is possible.

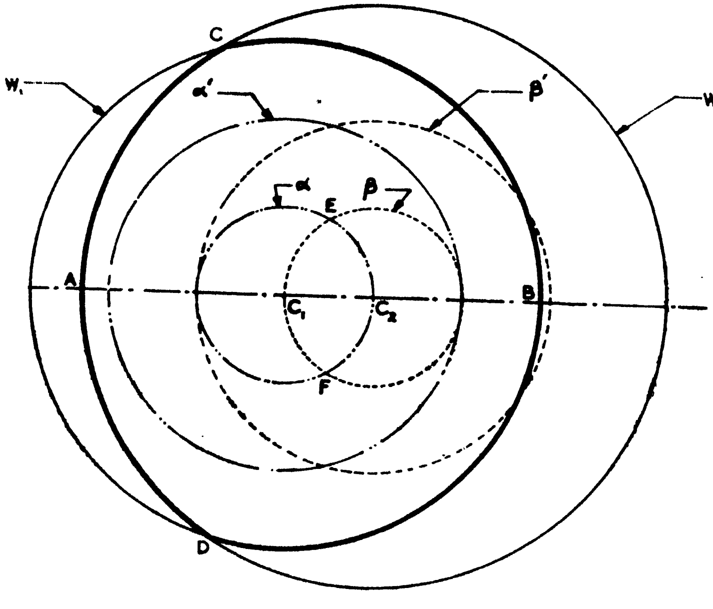


Figure 1.

This paper is concerned with the case when the two wavefronts are *identical*, a wavefront being tested against a "sheared sight of itself". It will be shown later that the existence of solutions is not confined to the case of wavefronts possessing revolution symmetry, but may also be extended to include simple astigmatism and coma.

When one is permitted to rotate one wavefront about a principal ray, in addition to being allowed lateral shear, then added information on the asphericities may be obtained. For example, with a relative rotation through π the comatic errors will appear doubled, and with a rotation through $\pi/2$ the astigmatic errors will appear doubled. The wavefront shearing interferometer to be described produces both lateral shear and tilt. Other interferometers have been devised which produce rotatory shear about a principal ray, in addition to lateral shear and tilt.

§ 2. PRINCIPLE OF THE INSTRUMENT

Essentially the interferometer consists of two plane dividing films, D and S, and two plane mirrors, M_L and M_R (figure 2). A convergent wavefront S_1, of which a principal ray meets D in B, is divided, and two identical coherent wavefronts emerge, one by reflexion in M_R and transmission through S (the R path) and one by reflexion in M_L and S (the L path), with principal rays along CE and CE′ respectively. An eye placed to receive these emergent wavefronts will see two apertures sheared with respect to one another (figure 2). The magnitude of the shear is continuously variable by rotation of S, the shear plate.

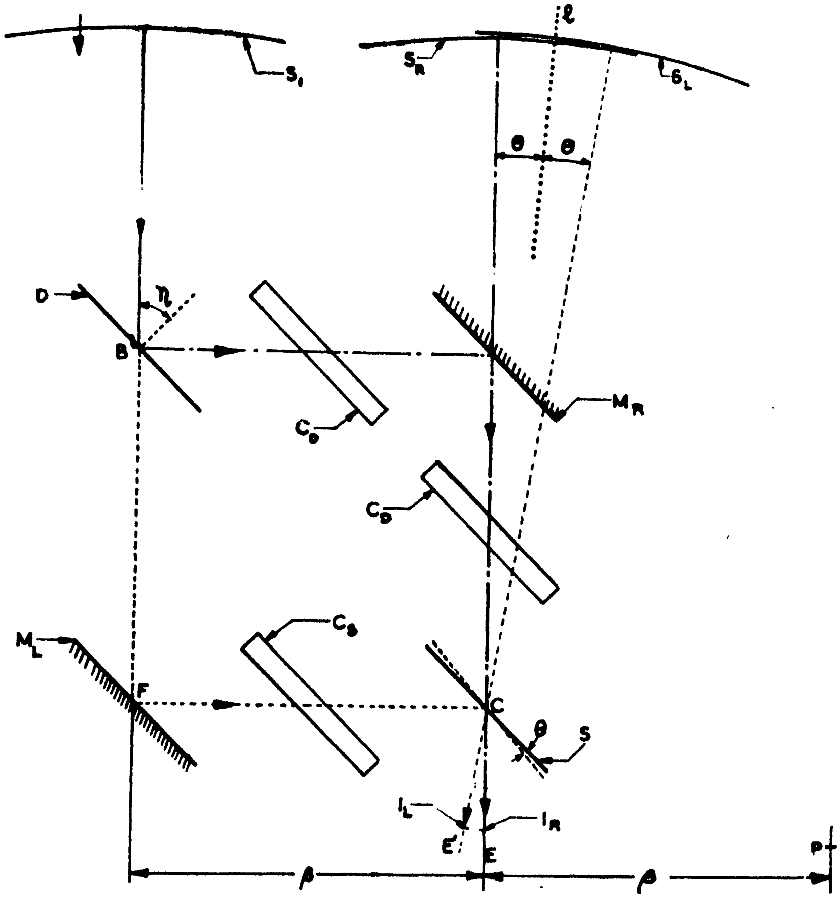


Figure 2.

The number of fringes and the tilt between the wavefronts will depend on the relative positions of the two images I_L and I_R.

Separations in the plane of the paper give fringes perpendicular to that plane and path length changes, while separations out of the plane of the paper give meridional section fringes. The angle between the emergent principal rays will be twice the difference between the angles of the mirror pairs D, M_R and S, M_L. If the latter two angles are made zero, then the apparatus consists of two parallel mirror pairs, and there is no shear between the emergent wavefronts. When the shear plate S is rotated through an angle θ about any point, the angle between CE and CE′ is 2θ. If the axis of rotation is chosen so that on shearing there is no relative positional change between I_L and I_R, then the number, tilt and whiteness of the fringes, will change only because of the wavefront's relative shear. This is so when I_L and I_R are located on the shear plate at C, and the latter is rotated about an axis in its plane passing through C during the shearing operation. Equalization of the path lengths may be effected exactly by a linear translation of M_L along BF, together with a rotation about F, so that the normal to M_L always bisects the angle BFC—a motion possible with a mechanical linkage. A sufficient approximation to this control, is a rotation of M_L about a point P twice as far from BF as from CE. If D and M_L remain a parallel pair, and are rotated as a whole about an axis parallel to CE, then meridional section fringes will be obtained by the separation of I_L and I_R out of the plane of the paper.

The interferometer is then a device which will interfere a wavefront with a sheared sight of itself. *Continuous* control of the shear, and of the meridional or sagittal fringes is effected and white light fringes are obtained.

Compensation

The dividing films must be supported on transparent plates of finite thickness, and when shear is present an automatic compensation of these thicknesses may be made. When the components of the interferometer are all parallel and the test wavefront is incident at an angle η on D (figure 2), there is no shear and by both the R and the L paths the wavefront encounters one plate at an angle η. If shear is obtained by rotating S through an angle θ, and a new principal ray R be defined, through the centre of the interferogram, then by the R path the wavefront will encounter S at an angle η, and by the L path it will encounter D at an angle $(\eta+\theta)$. To effect compensation of this difference, it is necessary to add two further plates, C_D and C_S, of the same thicknesses respectively as D and S. C_S is fixed parallel to the shear plate and rotates with it during the shearing operation, whilst the second plate, if placed at C_D, is made to rotate at twice the rate of the shear plate and in the opposite direction, and if placed at C'_D at twice the rate in the same direction. Exact compensation may thus be effected by a 1 : 2 gear ratio between S and C_D, or an equivalent mechanical linkage. The errors introduced by exact compensation are investigated in the appendix.

§ 3. EXPERIMENTAL

The apparatus in an experimental form is shown in plate 2. All the plates are in holders which are kinematically attached to their supports, and all possess altazimuth line-up adjustments. The plate D and mirror M_L (a parallel pair) can be rotated as a whole by small amounts by means of the screw T, thus obtaining the vertical image separation necessary for meridian section fringes. The path length control is the approximate one mentioned in § 2, working on screw control L.

and pivoting about the point P. The 1 : 2 gear ratio between the shear plate S, plus its compensator C_S, and the dividing plate compensator C_D, is anchored through a steel tape to two drums on the underside of the apparatus. The shear control is by means of the screw A working on a lever attached to the drum below C_D. Finally the apparatus is on a carriage, which can be moved in two directions at right-angles, in order to position the wavefront " focus " on the shear plate at its axis of rotation.

Departures from flatness of the plates and mirrors over the areas used, will appear as errors of the same order of magnitude in the interferogram interpretations. Relative variations in the plate thicknesses will also introduce spurious aberrations, for the magnitudes of the plate aberrations are proportional to their thicknesses, and if these are not correct then the compensation will be inexact. It is shown in the appendix that when, for instance, it is required to test an F.5 cone to an accuracy of 1/10 fringe, the plate thicknesses must be equal in pairs to about 1/30 mm. for plates 5 mm. thick. This tolerance will be proportional to the desired accuracy, and inversely proportional to the square of the test aperture. In a similar way errors in plate orientations lead to errors of compensation, and it is shown that for the same accuracy of measurement, the plate orientations must be correct to about 9′ of arc. Thus sufficiently exact compensation is not difficult.

Even so, when no compensating plates are used, the interferometer still has its uses. In this condition the path lengths in glass do not remain equal on shearing, and in addition astigmatic and comatic errors are introduced by the difference in plate orientations. Choosing the meridional focal setting, the fringes, though unequally spaced, will be straight in the absence of spherical or zonal aberration. The limit of accuracy of the instrument then depends on the spurious comatic error, and setting the maximum tolerable error at 1/100 fringe, it may be seen that it is possible in this way to test an F.7 cone sheared " edge over centre ".

A restricted source is necessary and the permissible size may easily be calculated. In the direction parallel to the plates, the source may be extended provided that the slit images are parallel to the axis of rotation of the shear plate. Notwithstanding this, there is no point in extending the images beyond the eye pupil diameter for visual observation. In the direction of shear, the source size is limited by the fact that during the shearing operation the images are rotated with respect to one another. The fringe visibility in a wavelength λ, with a slit width d, and shear angle θ (figure 2) may be calculated by considering the two wavefronts which leave the aperture as having been produced by two identical sources of width d rotated through an angle ψ with respect to each other, as shown in figure 2 (*a*). We are concerned with the phases of the waves from the sets of points A(≡A′), B(≡B′), . . ., and the intensity at a point defined by the parameter k is

$$I_k = \int_{d/2}^{d/2} \cos^2\left(\frac{2\pi x\psi}{\lambda} + k\right) dx$$

where x, as shown, is the distance from the intersection of the sources, and the limits of integration represent the whole width of the slit. Writing $\pi\psi/\lambda = a$ and $2ax + k = z$, the intensity becomes

$$I_k = \frac{1}{2a}\int_{k-ad}^{k+ad} \cos^2 z \,.\, dz = \frac{d}{2} + \frac{\lambda}{4\pi\psi}\cos 2k \sin\frac{2\pi\psi d}{\lambda},$$

so that the maxima and minima of I are

$$\frac{d}{2} \pm \frac{\lambda}{4\pi\psi}\sin\frac{2\pi\psi d}{\lambda}$$

and the visibility of the fringes

$$V = \frac{I_{\max} - I_{\min}}{I_{\max}} = \frac{\dfrac{\lambda}{2\pi\psi}\sin\dfrac{2\pi\psi d}{\lambda}}{\dfrac{d}{2} + \dfrac{\lambda}{4\pi\psi}\sin\dfrac{2\pi\psi d}{\lambda}}.$$

Now the shear angle θ is $\psi/2$, and if we write $4\pi\theta d/\lambda = \phi$, the expression for the visibility becomes

$$V = \frac{2\sin\phi}{\phi + \sin\phi}.$$

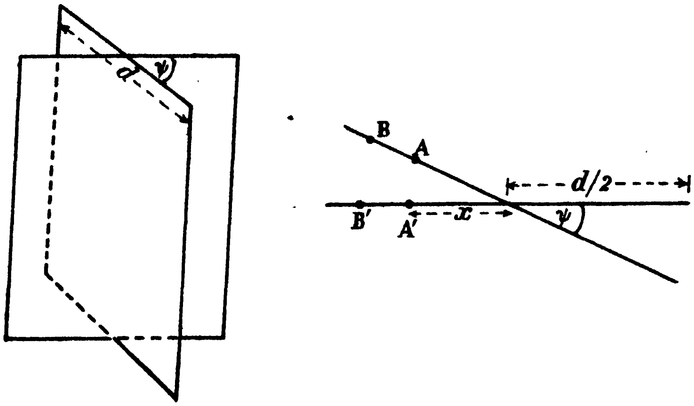


Figure 2 *a*.

For an F.10 cone sheared until the aperture edges are over the centres, and in a wavelength of 5000 A., this expression gives the permissible slit width for 60% visibility as $3{\cdot}25 \times 10^{-4}$ cm. This particular visibility is, as a matter of fact, about the value obtained with contact fringes between unsilvered plates, but much smaller values can be used satisfactorily.

The production of slits of the necessary width is not difficult, and one has been made which is adjustable from 2×10^{-2} to 1×10^{-4} inches with certainty. Diffraction around the sides of the slit causes no trouble.

§4. CHARACTERISTICS OF SHEARED WAVEFRONTS

When two wavefronts containing identical asphericities are exactly overlapped, a single bright fringe will cover the field of view, and if a tilt is applied between them then straight fringes parallel to the intersection of the wavefronts will result. On shearing, an interferogram representing the difference between the parts of the wavefront used will be obtained in the overlap region. The asphericity noted in this interferogram will not be in name the same as that in the original, but will depend on it in a precise way. As an example, consider the effect of shearing two wavefronts containing identical first order coma given by $h = k \, . \, x \, . \, (x^2 + y^2)$, where h is the wavefront asphericity measured along the ray at the positional coordinate (x, y) (figure 3). Then the fringe pattern in the overlap region, given by the superposition of

$$h_r = k \,.\, (x+\alpha) \,.\, [(x+\alpha)^2 + y^2]$$

and

$$h_l = k \,.\, (x-\alpha) \,.\, [(x-\alpha)^2 + y^2]$$

(where each wavefront is sheared a distance α along the x or shear axis), contains

$k \,.\, 4 \,.\, \alpha^2 \,.\, (x^2+y^2)$, change of focus;
$k \,.\, 2 \,.\, \alpha^2 \,.\, (x^2-y^2)$, astigmatism;
$k \,.\, 2 \,.\, \alpha^3$ change of path length.

In the same way it may be shown that pure spherical aberration shears into coma and a tilt, while pure astigmatism with axes along and perpendicular to the shear direction, produces in that direction a tilt which can be annulled by a lateral displacement of the apparatus. Astigmatism with axes skew to the shear direction produces, in addition, tilt perpendicular to the shear direction. Finally, when the wavefronts have revolution symmetry the interferogram represents an asphericity which is antisymmetrical about the centre O (figure 3).

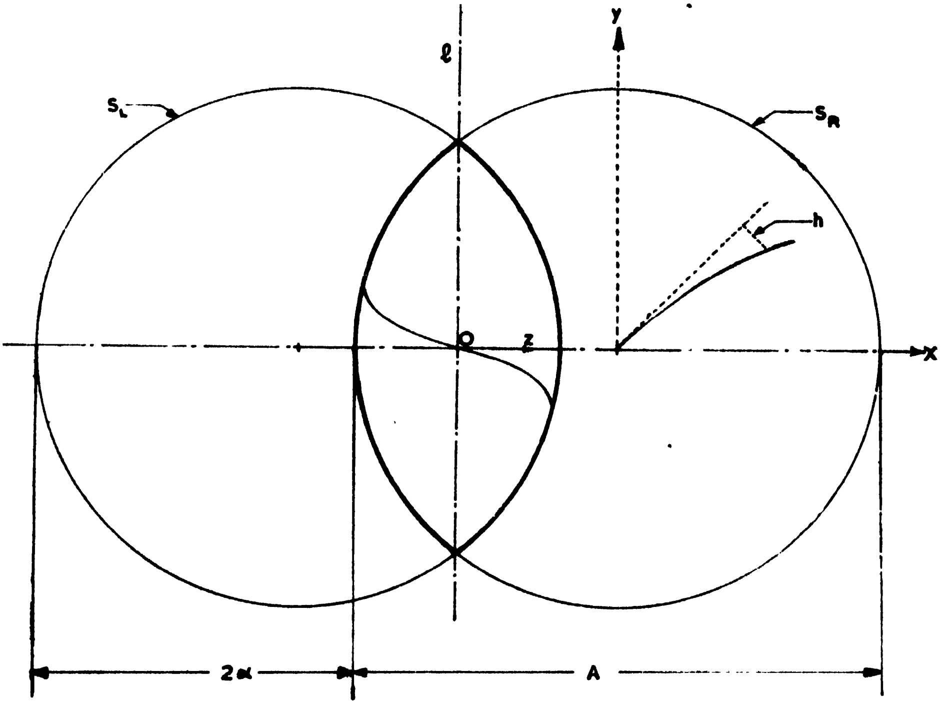


Figure 3.

The case of sheared paraboloids is of practical interest. The distance D between a paraboloid and a contacting sphere of the same polar radius of curvature ρ, measured along a diameter to that sphere at zonal radius z, is given by

$$D = \rho - \rho \,.\, \frac{[(1 - z^2/\rho^2)^{\frac{1}{2}} - (1 + z^2/\rho^2)^{\frac{1}{2}}]}{z^2/\rho^2}$$

$$= \frac{z^4}{8\rho^3} + \frac{7}{128} \cdot \frac{z^6}{\rho^5} + \ldots . \qquad \ldots\ldots(1)$$

Taking the leading fourth-power term only and shearing through plus and minus α as before, the interferogram retardation R is given by

$$R = \frac{z\alpha^3}{\rho^3} + \frac{z^3\alpha}{\rho^3}. \qquad \ldots\ldots(2)$$

Neglecting the term linear in z, the retardation across a meridian section is seen to be a cubic function of the position measured from the centre O (figure 3). By differentiating equation (2) with respect to α, it may be seen that the maximum amount of this cubic retardation occurs when the half-shear α is one-quarter of the linear aperture A. The maximum value of z will then be $A/4$, and the retardation at the interferogram edge is then $\frac{1}{256} \cdot \frac{A^4}{\rho^3}$. For a paraboloidal wavefront tested at its polar centre of curvature, this retardation in fringes will be $\frac{1}{2048} \cdot \frac{A}{F^3} \cdot \frac{1}{\lambda}$, where A is measured in the same units as the wavelength λ, and F is the focal ratio of the paraboloid. For an F.5 paraboloid of focal length 1 metre this gives about 1·6 fringes in a wavelength of 5000 A., and this retardation is of the same order as the fourth-power asphericity.

§ 5. CODA

Of the many methods of testing a wavefront, each has its particular advantages and limitations. The interferometer appears to have an advantage over ray methods in that it determines height differences directly rather than differences of slope, and a gain in simplicity over the usual methods of two-beam interferometry in that a separate comparison system is not required, Only experience will decide whether or not the advantages outweigh the concomitant limitations.

The technique is, however, applicable to any convergent wavefront. In its present form it would seem to have an astronomical application, in that it may be bolted to a telescope at the eyepoint to test the optical system under working conditions with a star as source. " Turned edge " is very easily seen in this way, as is also small magnitude rapid zonal error. Plate 2 shows the fringe pattern obtained under small shear from a nominally spherical speculum mirror which possessed a turned edge, with zonal error in addition to a scratched surface. Use may also be found for the instrument in the routine testing of an astronomical mirror during figuring operations.

§ 6. ACKNOWLEDGMENTS

The author wishes to express his thanks to Professor A. M. Tyndall, to Dr. C. R. Burch for unflagging interest and encouragement and to Mr. F. T. Bannister for assistance in the construction of the first instrument which excluded plasticene.

APPENDIX

Aberrations of skewed plane parallel plates

If a principal axis of a spherical wavefront, incident on a plane parallel plate of thickness t and refractive index μ, lies along the normal to the plate, then the longitudinal aberration S_ϕ of a ray at an angle ϕ to the plate normal is given by

$$S_\phi = \frac{t}{\mu}\left\{1 - \left[\frac{1-\sin^2\phi}{1-\sin^2\phi/\mu^2}\right]^{\frac{1}{2}}\right\}.$$

For a plate of refractive index 1·5 the magnitude of this aberration per unit plate thickness is given in table 1.

Table 1

ϕ°	S_ϕ/t
0	0
10	0·00568
20	0·02326
30	0·05429
40	0·10144
50	0·16824
60	0·25842
70	0·37413
80	0·51319
90	0·66667

In figure 4 the distance P_0P_ϕ is equal to S_ϕ where P_0 is the paraxial focus. Define now a new principal ray to be that ray at an angle ϕ to the plate normal, and let rays at $+\theta$ and $-\theta$ to it meet it in A_+ and A_- respectively.

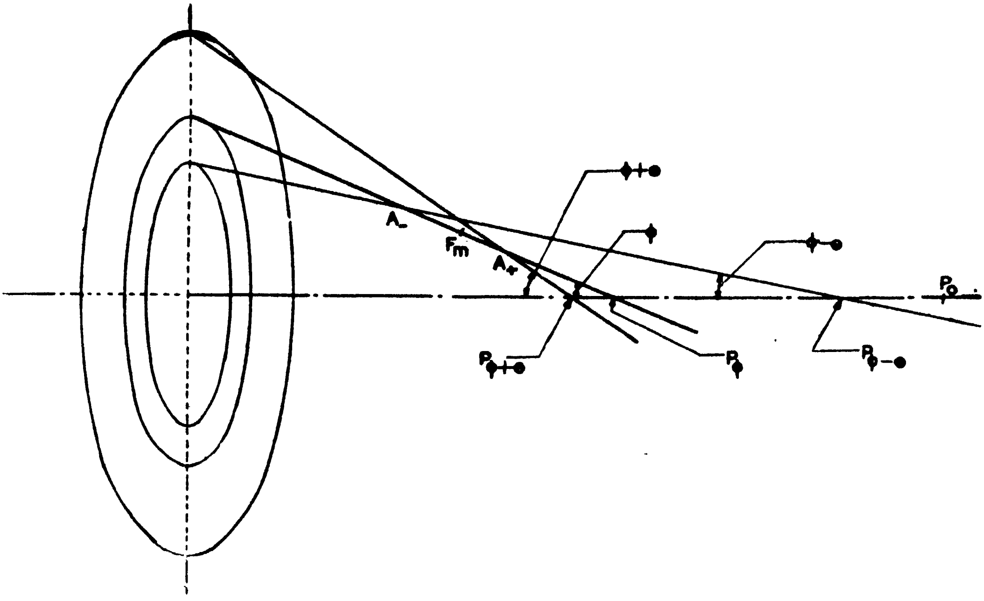


Figure 4.

If θ is reduced without limit the focus for rays paraxial with the new axis is determined at F_m, somewhere between A_+ and A_-. F_m is the location of the meridional focal line; the sagittal line, for fans out of the plane of the paper, is at P_ϕ. The distance F_mP_ϕ may be called the astigmatic interfocal distance.

With no restriction on θ the distance $\mathrm{A\_P}_\phi$ is given by

$$\mathrm{A\_P}_\phi = \sin(\phi-\theta)\,.\,\frac{[S_\phi - S_{\phi-\theta}]}{\sin\theta}$$

$$= [\sin\phi\,.\,\cot\theta - \cos\phi]\,.\left\{\theta\,.\,\frac{\partial S}{\partial\phi} - \frac{\theta^2}{\underline{|2}}\,.\,\frac{\partial^2 S}{\partial\phi^2} + \ldots. + (-1)^{n-1}\,.\,\frac{\theta^n}{\underline{|n}}\,.\,\frac{\partial^n S}{\partial\phi^n} + \ldots\right\}.$$

Expanding $\cot\theta$ in powers of θ this becomes

$$
\begin{aligned}
A\_P_\phi &= \sin\phi\,.\,\frac{\partial S}{\partial\phi}\\
&\quad + \theta\,.\left\{-\cos\phi\,.\,\frac{\partial S}{\partial\phi} - \frac{1}{2}\,.\,\sin\phi\,.\,\frac{\partial^2 S}{\partial\phi^2}\right\}\\
&\quad + \theta^2\,.\left\{-\frac{1}{3}\,.\,\sin\phi\,.\,\frac{\partial S}{\partial\phi} + \frac{1}{2}\,.\,\cos\phi\,.\,\frac{\partial^2 S}{\partial\phi^2} + \frac{1}{6}\,.\,\sin\phi\,.\,\frac{\partial^3 S}{\partial\phi^3}\right\}\\
&\quad + \theta^3\,.\left\{\frac{1}{6}\,.\,\sin\phi\,.\,\frac{\partial S^2}{\partial\phi^2} + \frac{1}{6}\,.\,\cos\phi\,.\,\frac{\partial^3 S}{\partial\phi^3} - \frac{1}{24}\,.\,\sin\phi\,.\,\frac{\partial^4 S}{\partial\phi^4}\right\}\\
&\quad + \ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}
\qquad \ldots\ldots(3)
$$

In this expression the term $\sin\phi\,.\,\partial S/\partial\phi$, which is independent of the aperture angle θ, gives the astigmatic interfocal distance $\mathrm{F}_m\mathrm{P}_\phi$.

These aberrations may be referred to the great sphere as follows for the purpose of error measurement in terms of fringes. The excess retardation of a ray at angle θ to the principal ray is given by

$$R_\theta - R_0 = \int_0^\theta L(\theta)\,.\,\sin\theta\,.\,d\theta, \qquad \ldots\ldots(4)$$

where $\mathrm{L}(\theta)$ is the longitudinal aberration measured along the principal ray, and is the distance $\mathrm{A\_P}_\phi$. Writing equation (3) as

$$A\_P_\phi = A + \alpha\,.\,\theta + \beta\,.\,\theta^2 + \gamma\,.\,\theta^3 + \ldots.,$$

then equation (4) becomes

$$
\begin{aligned}
R_\theta - R_0 &= \theta^2\,.\,\frac{A}{2} && \text{astigmatism;}\\
&\quad + \theta^3\,.\,\frac{\alpha}{3} && \text{primary coma;}\\
&\quad + \theta^4\,.\left(\frac{\beta}{4} - \frac{A}{24}\right) && \text{primary spherical aberration;}\\
&\quad + \theta^5\,.\left(\frac{\gamma}{5} - \frac{\alpha}{30}\right) && \text{secondary coma;}\\
&\quad + \ldots. && \text{higher terms.}
\end{aligned}
\qquad \ldots\ldots(5)
$$

The values of these coefficients for a plate thickness unity, and refractive index 1·5, around $\phi = 45°$ are given in table 2.

Table 2

ϕ°	A	$\alpha/3$	$\beta/4-A/24$	$\gamma/5-\alpha/30$
33	0·136	−0·120	0·0563	−0·0002
35	0·155	−0·129	0·0565	+0·0015
37	0·176	−0·138	0·0562	0·0038
39	0·196	−0·147	0·0566	0·0067
41	0·219	−0·156	0·0554	0·0100
43	0·244	−0·165	0·0528	0·0139
45	0·270	−0·174	0·0504	0·0186
47	0·300	−0·181	0·0470	0·0241
49	0·327	−0·189	0·0429	0·0305
51	0·357	−0·197	0·0399	0·0375
53	0·390	−0·204	0·0302	0·0454
55	0·423	−0·211	0·0220	0·0543

From this table the aberrations introduced into a cone of semi-angle θ, incident at an angle ϕ to a plane parallel plate, may be determined. For example, from the first term of equation (5) it will be seen that the astigmatism introduced into a cone incident at ϕ° to a plate of thickness t is given as $\frac{A}{2} \,.\, \theta^2 \,.\, \frac{t}{\lambda}$ fringes, in a wavelength of λ. For an incidence of 45°, $A=0{\cdot}27$ (table 2), and with an F.5 cone, plate thickness 0·5 cm., and wavelength 5000 A., this astigmatism is about 13·5 fringes. Thus the compensating plates of the interferometer must be equal in thickness to the dividing plates to about 1/27 mm., in order to test such wavefronts to an accuracy of 1/10 fringe. With the same accuracy of thickness the second term of equation (5) shows that the associated primary coma is of the order of 1/100 fringe.

Consideration must also be given to errors in orientation of the interferometer plates. A change in plate orientation from 45° to 43° produces a change in the coefficient of A of 0·026 or a reduction of about 1·3 in the number of fringes of astigmatism. To test to an accuracy of 1/10 fringe will therefore require the plate orientations to be correct to about 9′ of arc.

The above tolerances are smaller than is actually necessary in the wavefront shearing interferometer, since the angular opening of the interference aperture is of necessity always less than that of the incident wavefront. Particularly is this so for large values of the shear, when also the coefficients of table 2 are larger.

From equation (3) it would appear that the higher order meridional aberrations do not increase excessively, but a rigorous proof of this has not been attempted. It will also be appreciated that the calculations consider only the meridian section of the wavefront.

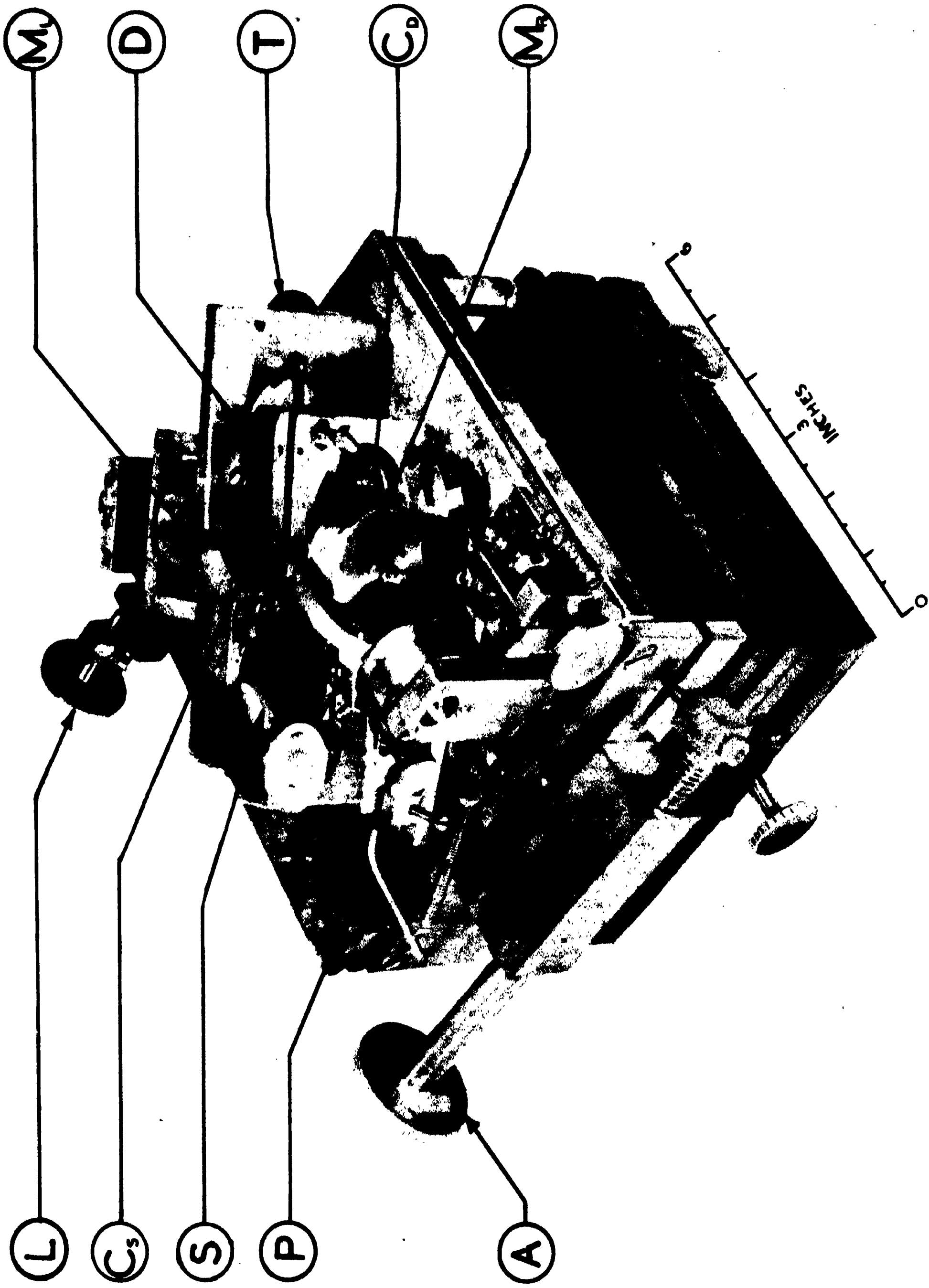
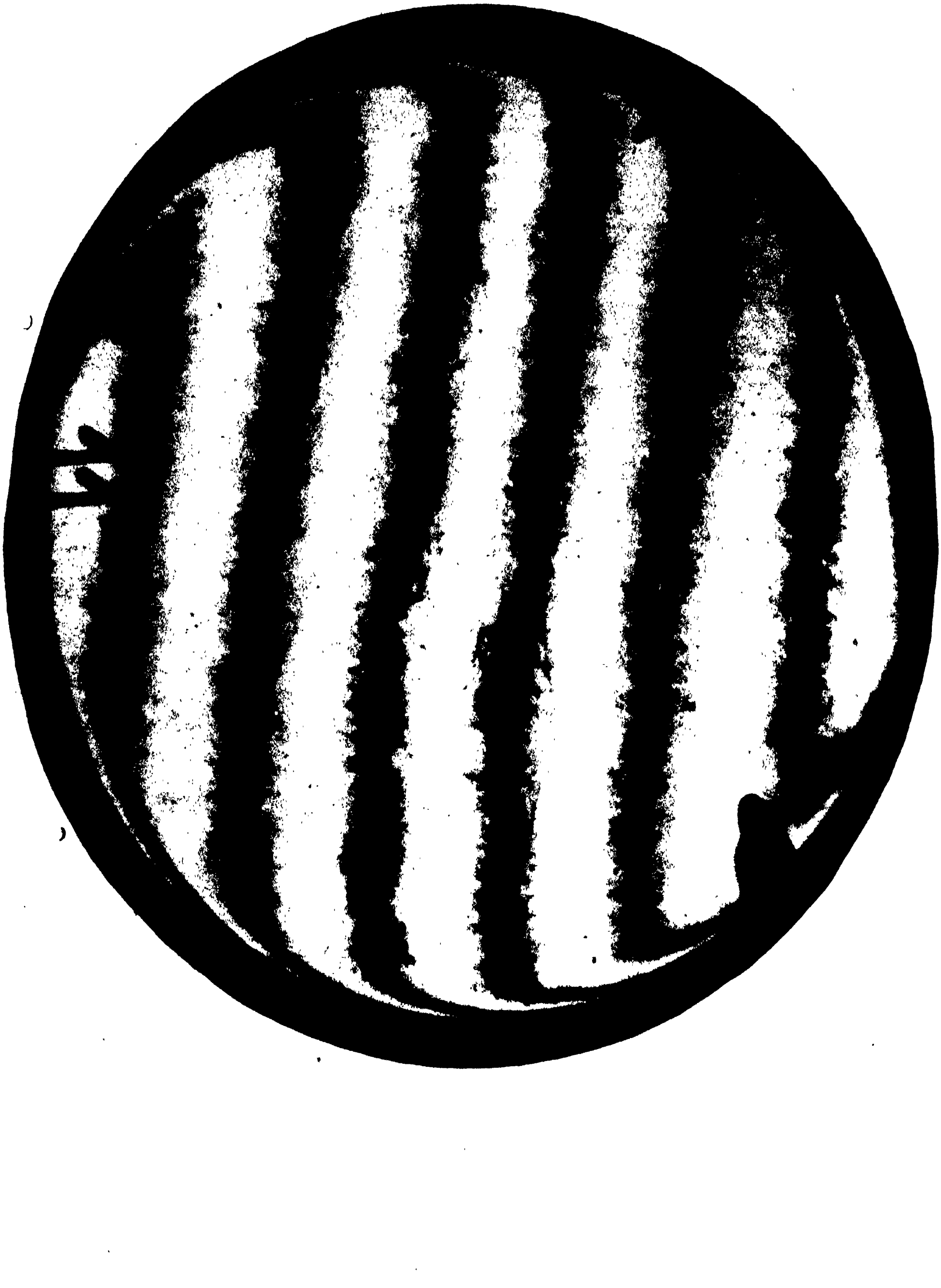


Plate 1.



A TRANSPARENT-REPLICA TECHNIQUE FOR INTERFEROMETRY

By R. C. FAUST and S. TOLANSKY,
Royal Holloway College

MS. received 15 *May* 1947

ABSTRACT. A description is given of a transparent-replica technique which allows the surface of an opaque body to be examined interferometrically using the transmitted multiple-beam fringe pattern. This avoids the difficulties inherent in the technique of reflected fringes. A replica made from methyl methacrylate polymer is found to reproduce features both in extension and in depth to within close limits. Contours of only 40 A. in height are faithfully reproduced, and it is considered that a 10 A. change would be copied. The replica shows an overall shrinkage of the order of ½ per cent, but this is not a serious drawback.

The technique is tested on glass, mica and calcite, and is then applied to the examination of a coarsely polished metal surface, revealing features of interest. The new technique opens up further possibilities in the application of multiple-beam interference studies to the examination of the surfaces of polished metals. In particular, phase-effect errors are eliminated entirely.

§ 1. THE NEED FOR A REPLICA TECHNIQUE

RECENT developments in precision multiple-beam interferometry (Tolansky, 1946 a) for the study of surface topography are such that it has been found desirable to evolve a reliable replica technique in accordance with the following requirements.

The examination of an almost flat surface of a transparent body (e.g. a clear crystal, or a thin film) by multiple-beam interferometry is relatively simple, and fine sharp precision fringes are readily obtained (Tolansky, 1946 a) if the two surfaces concerned are suitably silvered and maintained at a separation no greater than a few wave-lengths of visible light. If a study is to be made of an appreciably concave surface, or of features lying below the general level of the specimen (for example deep etch pits), then the condition of close approach cannot be realized. Clearly, if a reliable negative replica cast can be prepared, the originally depressed features will now be elevated and can therefore be brought to within the requisite distance from the reference flat.

Of greater value is the application of a reliable replica technique to the examination of opaque surfaces, such as those of metals, as the following drawbacks are associated with the direct interferometric study of a metallic surface:—

(1) Because of opacity a back-reflection technique must of necessity be employed (Tolansky, 1945). If a low-power objective is in use, the working distance is sufficient to allow the insertion of a 45-degree reflector between the objective and the optical flat and no great difficulties are involved. This method is only of practical value for useful magnifications not exceeding × 50. If higher magnifications are required, the surface illumination practised in standard metallurgical microscopes must be resorted to. Here the reflector is between

the objective and the eyepiece, the objective performing the dual rôle of condensing light on to the specimen and at the same time forming an image of the illuminated surface. However, for multiple-beam interferometry the light falling on the specimen must be a strictly parallel beam at normal incidence. Hence the microscope objective must simultaneously produce a parallel beam from a small image formed at its rear focus and also act in its normal capacity as an image-forming lens. To carry out both functions correctly it would appear that a specially designed lens is necessary. The attainment of useful magnifications much in excess of × 50 is therefore associated with considerable experimental difficulties.

(2) In a forthcoming publication (Tolansky, 1947) it is shown that the visibility of the reflected fringe system is critically dependent upon the light-absorption of the silver film deposited upon the reference flat. If effective reflectivities as high as those permissible with transmission fringes are employed with the back-reflection technique, the fringe visibility is poor, and consequently the fringes are hard to detect. The visibility can be improved by lowering the reflectivity of the silver film on the flat, but this leads to an increased fringe width.

(3) The fact that the reflected fringes are fine dark lines on a broad bright background necessitates the use of monochromatic radiation. This is a drawback since, as has been shown before, considerable advantage in interpretation arises from the use of a mixed group of a small number of distinct wave-lengths. This restriction to monochromatic light means that a series of separate photographs must often be taken, each with a different wave-length, thus increasing both the labour and the difficulty of interpretation.

(4) Because of the above restriction the valuable "crossed-fringe" technique (Tolansky and Wilcock, 1946) cannot be adopted.

(5) A principal difficulty which can lead to serious major errors arises when metal surfaces, particularly alloys, are under examination. The point at issue is the question of differential phase change at reflection. Suppose the surface to possess a coarse, heterogeneous structure. Such a structure could arise for instance from any of the following: (*a*) different alloy constituents, (*b*) differential local ageing, (*c*) local polishing differences, (*d*) corrosion, (*e*) film formation. If the phase change at reflection varies over such local features then a serious error arises in interpretation since an optical change in phase can then be misinterpreted as a considerable metrical change in level. There would appear to be two ways of avoiding this error. One method would be to evaporate a fairly thick film of silver over the metal surface, and thus impose a uniform phase change (that for the silver) over the whole area. This appears attractive, but there may also be an error involved in so far as we are ignorant as to whether regional variations beneath the silver (e.g. corrosions) locally affect the phase change. (It is recalled that the silver film will be less than 1000 A. thick.) The second method for overcoming this difficulty is to develop a technique for making a reliable transparent replica casting, which eliminates phase effects.

Thus it is seen that if a transparent replica technique can be evolved, all the difficulties discussed will be removed and the simpler methods available

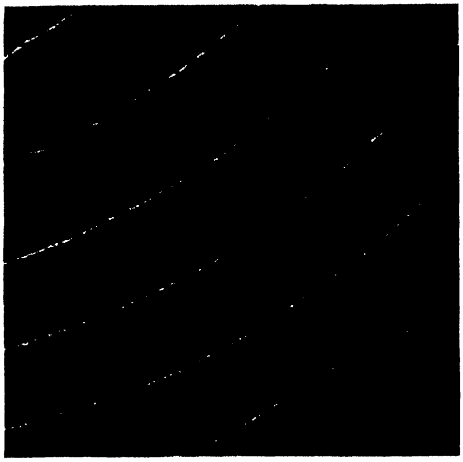


Figure 1.

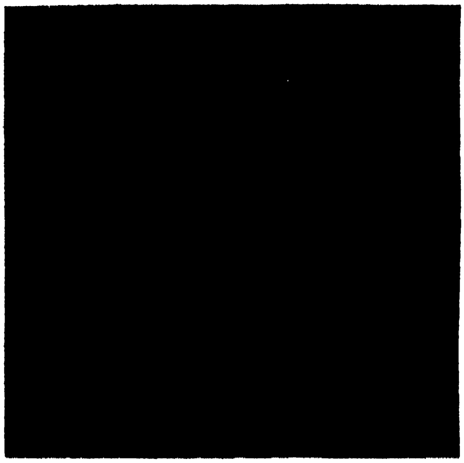


Figure 2.

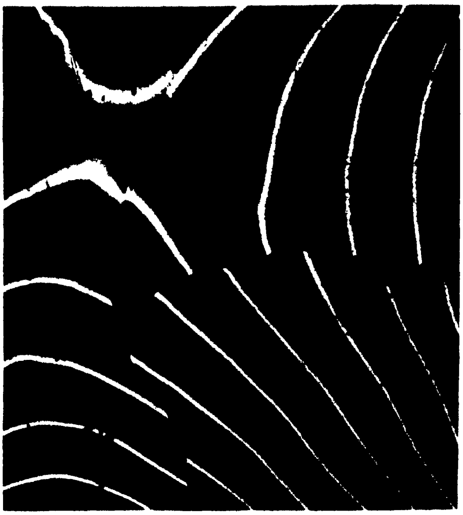


Figure 3.

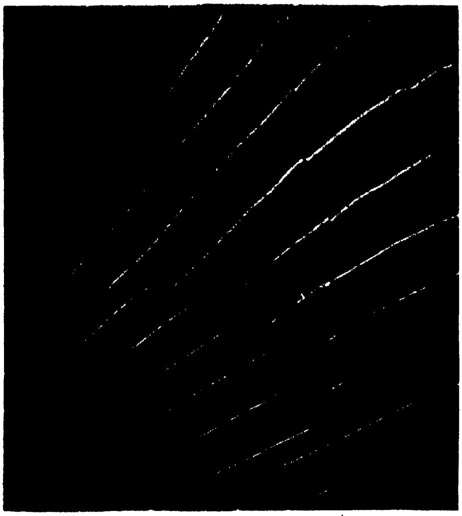


Figure 4.

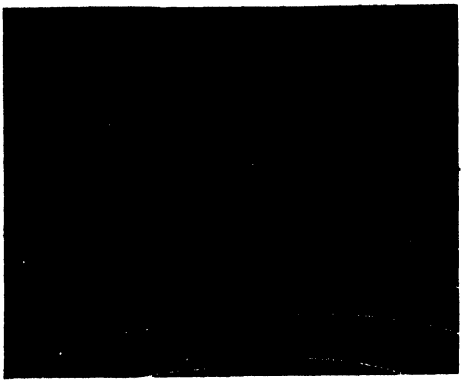


Figure 5.

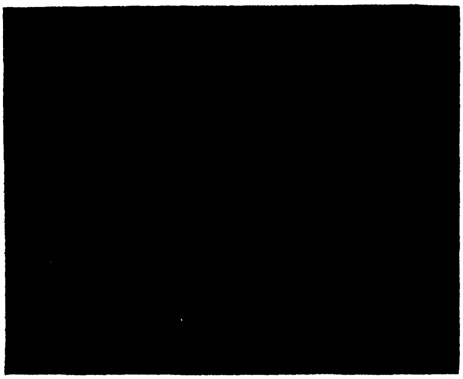


Figure 6.

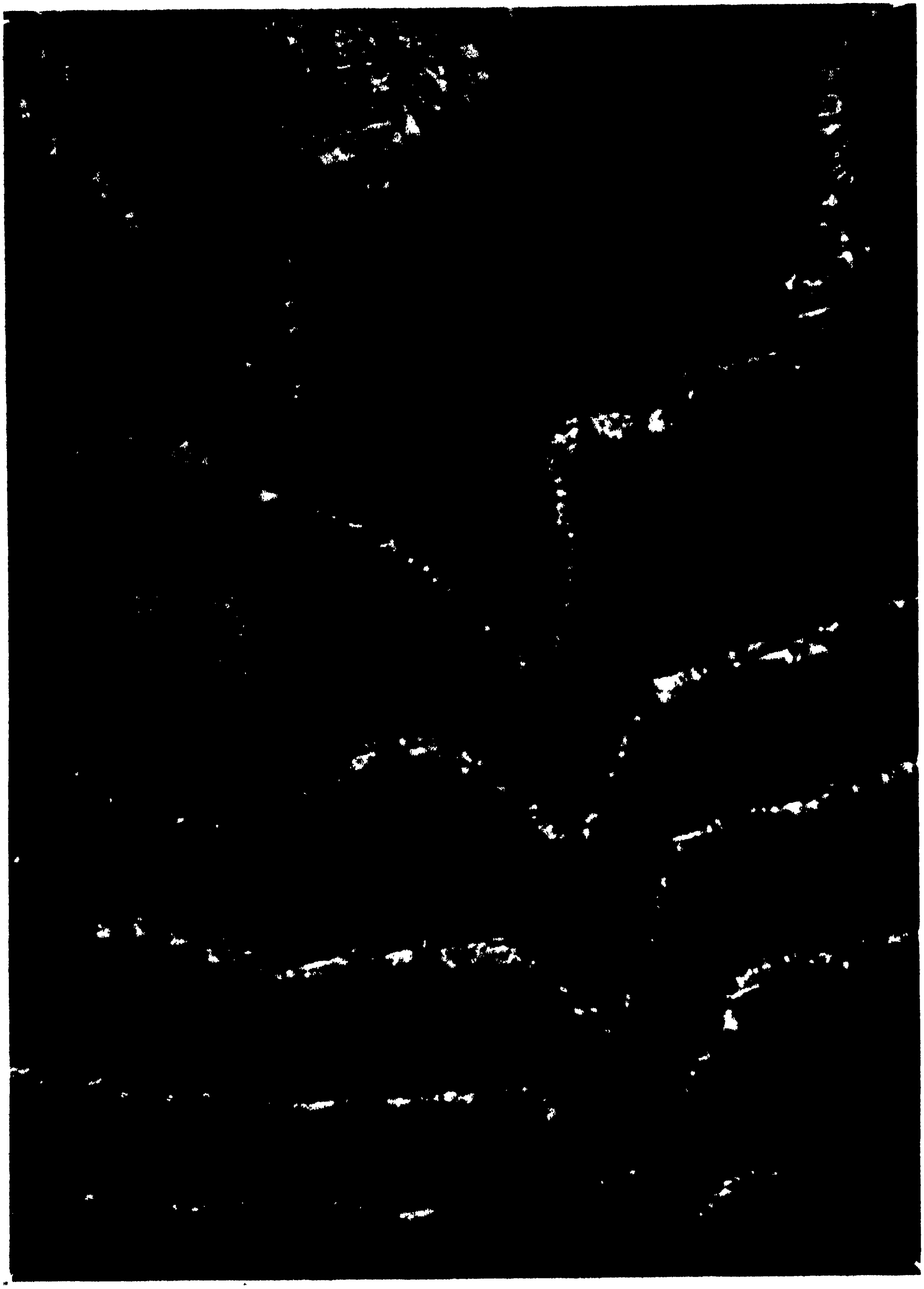


Figure 7.

for transparent bodies can be applied to the study of the surfaces of opaque solids. The realization of this possibility would be a considerable advantage.

Such a technique has been successfully developed, and this paper is largely devoted to a report on the reliability and faithfulness of reproduction realized by the procedure.

§ 2. THE REPLICA TECHNIQUE

A replica is required with the following characteristics:

1. It must be transparent.
2. Features in extension (i.e. across the surface), the dimensions of which are at least as small as 1/500 millimetre, must be reproduced with high fidelity.
3. Features in depth must be reproducible to within molecular dimensions.
4. As the replica is to be handled for adjustment it should be robust.
5. It should be capable of remaining in a vacuum without distorting.
6. A copious gas stream must not be liberated from the replica when in a vacuum, otherwise a silver coating of low light absorption cannot be deposited upon the surface.

Two materials which it was considered might meet these requirements were the I.C.I. plastic products Transpex I (unplasticized polymethyl methacrylate) and Transpex II (unplasticized polystyrene). Of these the methacrylate polymer is much to be preferred since it is far more readily degassed in the vacuum than the polystyrene, and all the measurements and photographs presented in this paper were made with Transpex I replicas. The material employed is in sheets about 4 mm. thick. From such a sheet a small piece is cut and washed with soap and warm water and then transferred to an oven where it is maintained at a temperature of about 170° C. At this temperature the plastic rapidly dries and is brought to a state suitable for moulding. Drying with a fabric is not attempted since frictional charges are readily built up, resulting in the collection of fibres and dust. The specimen, mounted in a suitable cement, is placed on an electric hot plate and brought to a temperature of 140° C. The plastic is placed on the specimen and then covered with a piece of hot plate glass. Upon the application of slight pressure the soft plastic flows and moulds itself on the one hand to the specimen surface and to the plate glass surface on the other. The plate glass is depressed until it finally rests on an accurately machined brass ring of constant height, this ring being in turn supported on a plane containing the specimen face. Consequently the finished replica approximates closely to a plane parallel sheet. This is desirable as the presence of an appreciable wedge-angle between the faces of the replica gives rise to ghost images and secondary fringes. The whole is allowed to cool slowly (at least an hour) to room temperature to avoid the setting up of internal strains. The replica and specimen are easily separated, the former then being coated with an evaporated silver film. Due to the flow characteristics of the plastic, the moulding pressure employed is quite small and it is not considered that specimen deformation is introduced.

The specimen is mounted, according to its character, in a suitable cement capable of withstanding the moulding temperature. For this purpose an artificial stone, Kaffir "D", has been found generally useful. An alternative investment is a phenol-formaldehyde thermosetting resin. Prior to the preparation of a replica it is helpful to clean the specimen by employing the well-known method of repeatedly stripping off collodion films.

§ 3. REPRODUCTION IN EXTENSION

To test the reliability of the reproduction in extension, casts were made from plane gratings ruled on metal. The replica gratings (which are quite robust and far superior to the usual Thorpe replicas found in teaching laboratories) were mounted on a spectroscope and the number of lines per inch determined with the sodium D lines in the usual manner. The values so obtained are compared below with the data marked on the original gratings.

	Lines per inch		
Original	2400	14200	17300
Replica	2412	14270	17370
Difference	+0·5%	+0·5%	+0·4%

Of particular interest is the fact that the 17,300-line metal grating showed symmetrical ghost lines, whose intensity increased with order number. Identical ghosts were observed with the replica grating. Furthermore, the definition and resolution of the replicas were in each case identical with those of the parent grating. It is clear that in these three cases the lateral structure of the specimen surface was very closely reproduced by the replica, the whole scale of the structure having undergone, however, a contraction of about ½ per cent.

The ½ per cent contraction arises from the difference between the thermal expansion of the metal and that of the plastic. The plastic tends to conform to the specimen structure down to the setting point (about 100° c.), at which state it will possess the structure of the thermally expanded metal at that temperature. As the coefficient of thermal expansion of the metal is about $0{\cdot}2\times10^{-4}$ per degree centigrade over the range 100° to 20° c., it contracts by about 0·2 per cent. The expansion coefficient of the plastic is higher, being about 1×10^{-4} per degree centigrade; consequently the plastic contraction is some 0·8 per cent. The difference between these two contractions, 0·6 per cent, represents the net shrinkage of the cooled plastic relative to the final cold state of the specimen. This figure is in close enough agreement with the measured values for the effective contraction. It follows that a sufficiently accurate correction can be made if 1×10^{-4} per degree centigrade be taken as a mean linear coefficient for the plastic. In the majority of cases this small correction can be disregarded.

§ 4. REPRODUCTION IN DEPTH

Whilst the many well-known replica techniques developed for electron microscopy indicate that faithful reproduction in extension might have been anticipated, such experience offers little evidence for the expected behaviour of a replica in terms of depth. It is this aspect which is the crucial one for interferometry, as it is in the determination of small heights and depths that the multiple-beam interference technique is so specifically powerful. It will now be shown that the reproduction in depth is very close indeed, and sufficient for many purposes. Replicas have been made of surfaces, the characteristics of which were already known from previous interference experiments, and the replicas then compared with the originals.

1. *Glass.* Figure 1 shows the multiple beam Fizeau fringes (λ 5461) formed when a replica of a piece of thin glass has been silvered and matched against a similarly silvered glass flat. The magnification is × 40. The fringes are typically those shown by glass. It has already been demonstrated elsewhere that the fine disrupted structure of the fringes is due to polish marks on the reference flat, this structure appearing only under critical illumination conditions.

2. *Mica.* It has previously been established that the steps appearing on the cleavage faces of muscovite are often small integral multiples of the 20-A. lattice spacing. A sample of muscovite was baked at 120° C. to drive off occluded water and a replica was then taken from a freshly cleaved face. Both the original and the copy were silvered and examined using Fizeau fringes. The characteristic cleavage steps were reproduced on the plastic, these steps being compared with the corresponding ones on the mica. It was known that a given cleavage step on mica is often constant in height to within a single molecular lattice over lengths of several millimetres. Measurements on the replicas showed that in this case also the step heights were true to within 5 A., the error in observation. The replica is so exact that it is quite easy to identify cleavage lines, and a precise comparison between the step heights on the mica and on the replica can be made. The following table is selected arbitrarily from such measurements, the step heights being in Ångström units. The experimental errors in each of the values quoted is of the order of 5 A.

Mica	385	505	1026	1280	1318	3237	4865
Replica	385	499	1020	1274	1302	3196	4822
Difference (%)		−1·2	−0·6	−0·5	−1·2	−1·3	−0·9

It is seen that over an extensive range the replica steps closely follow the original mica ones, but are consistently smaller than the latter by about 1 per cent. This shrinkage is not detected in the first step, being masked by the somewhat larger experimental errors. This 1 per cent shrinkage in depth is consistent with the shrinkage found in extension with the metal gratings, since the expansion coefficient of mica is considerably less than that of a metal.

Figures 3 and 4 show Fizeau fringes given by a sample of mica and by its replica. Figure 2 shows the same replica illuminated with the mercury yellow doublet instead of the green line (× 50). As the fringe pattern is determined by the relative disposition between the silvered surfaces of the specimen and the reference flat, it is natural that the general contours from the mica and from the replica should differ. The important points for comparison are:

(*a*) The correspondence in the outlines of the cleavage steps.

(*b*) The heights of the corresponding steps, which are not influenced by flexure of the specimen. (The replica photographs are mirror images of the original mica, and raised features on the latter become depressions on the former.) The scratch marks in figure 2 are on the unsilvered face of the thin mica slip and are, therefore, not reproduced by the replica.

The fidelity of reproduction is very clearly shown not only by the numerical data but also by the clearly defined small step running in a nearly vertical direction in the right halves of figures 2 and 4. This step, which is only about $\lambda/40$ high, is nevertheless quite readily detectable. Evidence concerning the lower limit of reproducibility is provided by an interesting feature which has emerged from the examination of one particular cleavage line. Upon running along the length of this step on the mica a discontinuity in step height was found to occur, the height changing from 1320 to 1279 A. (Each of these values is perhaps in error by 3 A.) This change of 41 ± 6 A. corresponds to two molecular lattices. The striking fact is that at this point the replica exhibited a corresponding change of 34 ± 6 A. Thus notwithstanding the length of the long flexible polymer molecules, the flowing plastic contours the mica so critically that a change in height of only two mica lattices is almost exactly followed.

Figure 5 shows fringes given by another mica replica (× 50). The apparent black gaps between major "plane" areas arise from the occurrence of several close, narrow, descending steps, which result in a loss of light. The characteristic smooth continuuity of the mica fringes, formerly shown to indicate that mica cleaves true to a molecular plane over extensive areas, also appears on the replica. Indeed, despite experience and familiarity extending over several years in the examination of the fringes given by mica, we are quite unable to distinguish between fringes given by micas and those given by replicas, confusing both types completely if descriptive marks are obliterated.

These observations indicate that topographical features, even perhaps as small as 10 A. in height, will be reproduced with high fidelity.

3. *Calcite.* Figure 6 shows the fringes given by a replica taken from a cleavage face of a calcite crystal (× 50). This photograph is characteristic, closely simulating the calcite fringes already familiar to workers in this laboratory. It was not considered profitable to make measurements on calcite, since it has been established (Tolansky and Khamsavi, 1946) that the application of light pressure to a calcite cleavage surface leads to the gliding of crystal units, and this changes the height of certain cleavage steps. The reproduction is included to show that a mould can be taken of a relatively soft, friable surface.

4. *Polished steel.* Figure 7 is a reproduction of the fringes (× 70) given by a replica from a piece of steel which had been crudely polished and then buffed. The region selected includes a scratch mark about half a wave deep. It can be assumed from the preceding observations that the replica faithfully reproduces the metal surface topography, and this being the case, a number of points of interest emerge.

The considerable fringe width is evidence of surface irregularities, all the bright points in a given disrupted and broadened fringe being equidistant from the reference surface. The fringe broadening is greatest when the fringe direction is parallel to the clearly delineated polish scratches. The maximum fringe width in this particular photograph is about one fifth of an order, corresponding to polish marks having a height (depth) of the order of $\lambda/10$. It is obvious that sharp fringes can only be formed with highly polished surfaces.

A clearly distinguishable feature on this interferogram is the well-marked ridge. A ridge on the replica corresponds to a rut, or scratch on the original.

One can clearly discern that the metal is ploughed up a small fraction of a light wave on either side of the scratch mark, which is little more than 1 mm. long and perhaps 2000 A. deep. It is to be emphasized that these features recorded here are quite free from differential phase-change errors.

§ 5. CONCLUSION

The replica technique as described has been subjected to critical tests on transparent materials, whose topographical features are already well established, in order to obtain a measure of its reliability. These experiments demonstrate that the technique employed is suited to the purpose of examining the topography of opaque bodies, such as metals. The manifold requirements demanded for a faithful replica technique are met. Reproduction in extension is correct to well within the desired limits, but a general shrinkage of some ½ per cent in the whole contour takes place when a replica is taken from metal. For many studies such an effect is of little consequence. In depth reproduction it is surprisingly good. Features whose contours are smaller than 300 A. reproduce with an accuracy which is within the experimental error of 5 A., and a change in height of only 40 A. has been copied with fidelity. It is probable that a change of only 10 A. would be followed. The shrinkage effect takes place in depth as well as in extension and can be easily corrected for, if desired.

The methyl methacrylate used for the replica takes a high grade silver coating, exhibiting low absorption, from which it can be concluded that the final evolution of gas from the plastic surface must be small. The softness of the plastic permits it to be brought into intimate contact with a high grade optical flat, a somewhat risky procedure with the hard metal. This intimate contact leads to improved fringe definition, according to views already developed elsewhere.

Although the existence or otherwise of any creep effects in the replica has not been intensively studied, their influence, if any, must be slight, since observation has shown that a replica grating has remained unaltered over several months.

§ 6. ACKNOWLEDGMENT

This investigation was made possible with the aid of a maintenance grant from D.S.I.R. to R. C. Faust.

REFERENCES

TOLANSKY, 1945. *Proc. Roy. Soc.*, A, **184**, 41, 51.
TOLANSKY, 1946 a. *Proc. Roy. Soc.*, A, **186**, 261.
TOLANSKY, 1946 b. *Proc. Phys. Soc.*, **58**, 654.
TOLANSKY, 1947. *Proc. Zeeman Congress, Amsterdam,* 1946 (in the press).
TOLANSKY and KHAMSAVI, 1946. *Nature, Lond.*, **157**, 661.
TOLANSKY and WILCOCK, 1945. *Nature, Lond.*, **157**, 583.

THE THEORY OF AN OSCILLATOR COUPLED TO A LONG FEEDER, WITH APPLICATIONS TO EXPERIMENTAL RESULTS FOR THE MAGNETRON

By C. DOMB,
Pembroke College, Cambridge
(sometime of Admiralty Signal Establishment)

MS. received 17 *December* 1946

§ 1. INTRODUCTION

THE phenomena associated with an oscillator having coupled circuits in its output were dealt with many years ago in the early papers of van der Pol (1922) and others. It was there shown how under certain conditions two stable frequencies were possible, and the oscillator might start up on either. The purpose of the present paper is to deal with the case when the output consists of a length of feeder with a mismatch at the end. A good deal of attention has been focused on this problem recently because of the centimetre wave technique in which no R.F. amplification is possible, and feeders are often some hundreds of wavelengths long.

The problem falls naturally into two categories:

(*a*) When the feeder is sufficiently short for the time of forward and backward travel of the wave along it to be small compared with the time of rise of the oscillator; this may be termed the Static Case.

The feeder may be considered as an ordinary impedance element and the problem bears many similarities to the case of coupled circuits.

(*b*) When the feeder is sufficiently long for the propagation time to be of importance; this may be termed the Dynamic Case. The form of the initial frequency modulation before the oscillator settles down to its ultimate frequency is now of great interest, particularly in the case of pulsed transmissions. This is fundamentally different from anything in the classical theory.

The theory was developed for a negative resistance oscillator with a general characteristic. Experimental data were available for the magnetron, and an attempt was made to apply the theory using an empirical characteristic. The normal method of doing this would be to make detailed impedance measurements throughout the output circuit and hence determine the power output as a function of load. It is shown, however, how the static frequency pulling and power output curves for lengths of line varying in phase from 0 to 2π can be used to provide the data required.

§ 2. THE STATIC CASE

It is well known that many of the properties of oscillators can be accounted for by replaçing the source of oscillations by a negative resistance. By this means we can deduce the frequency of the oscillations, but in order to determine the power output we must proceed to a further approximation. The condition for oscillation to be possible at a stable amplitude level is that the amount of power provided by the negative resistance must equal the power dissipated in the positive resistances in the circuit. If this source of power for any given amplitude level is greater than the power dissipated, the amplitude of the oscillations will increase. The simple assumption of a negative resistance is not sufficient to enable us to tell to what height the amplitude of the oscillations will rise. In order to do this we must assume a negative resistance characteristic—that is, a negative resistance which is a function of the voltage across it and which decreases when the voltage rises above a certain level. Whereas a simple negative resistance corresponds to a relation between current and voltage of the form $i = -av$, what is required is a relation $i = \chi(v)$ where $\chi(v)$ is of the form $i_0 - av$ for small v but flattens out as v increases. An example of this type is shown in figure 1. There will be one point on the curve at which the power provided by the source during a cycle of oscillation is exactly balanced by the power dissipated, and this is the height to which the amplitude of the oscillations will rise.

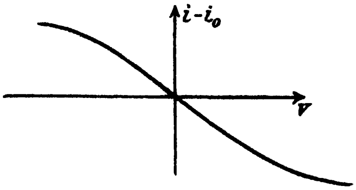


Figure 1.

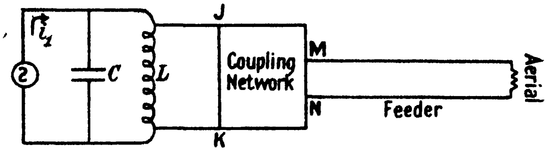


Figure 2.

Frequency and amplitude of a steady solution

The system with which we shall be concerned is shown diagrammatically in figure 2. However, it is not difficult to deduce the amplitude and frequency of steady sinusoidal oscillations for any general linear output network. Let v be the voltage across the generator. Then the differential equations for any network can be put in the form

$$\left.\begin{aligned} v &= z_{11}i_1 + z_{12}i_2 + \dots + z_{1n}i_n \\ 0 &= z_{21}i_1 + z_{22}i_2 + \dots + z_{2n}i_n \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ 0 &= z_{n1}i_1 + \dots \qquad + z_{nn}i_n \end{aligned}\right\} \qquad \dots\dots(1)$$

where z_{ij} is a function of the operator $p \equiv d/dt$.

Solving for i, in terms of v, we obtain

$$p^m\Delta(p)i_1 = p^m\Delta_1(p)v, \qquad \dots\dots(2)$$

where

$$\Delta(p) = \begin{vmatrix} z_{11} \dots z_{1n} \\ z_{21} \dots z_{2n} \\ \cdot\ \ \cdot\ \ \cdot\ \ \cdot \\ z_{n1} \dots z_{nn} \end{vmatrix} \qquad \Delta_1(p) = \frac{\partial\Delta}{\partial z_{11}} = \begin{vmatrix} z_{22} \dots z_{2n} \\ \cdot\ \ \cdot\ \ \cdot\ \ \cdot \\ z_{n2} \dots z_{nn} \end{vmatrix}$$

and the factor p^m is inserted to eliminate any inverse power of p.

Also, as mentioned above, we assume a relation

$$i_1 = -\chi(v)^*. \qquad \ldots\ldots(3)$$

Therefore

$$p^m\Delta(p)\chi(v) + p^m\Delta_1(p)v = 0. \qquad \ldots\ldots(4)$$

Equation (4) is a non-linear equation many examples of which have been considered by Appleton (1923), van der Pol (1934) and others. We shall be concerned only with cases in which a solution exists which is approximately sinusoidal. Assume a solution

$$v = A\cos\omega t = \mathrm{Re}\,(Ae^{i\omega t}). \qquad \ldots\ldots(5)$$

To deal with the non-linear terms of (4), i.e. those involved in $\chi(v)$, we follow a procedure similar to that of van der Pol. We assume that $\chi(A\cos\omega t)$ can be expanded as a Fourier Series,

$$\chi(A\cos\omega t) = a_0 + a_1\cos\omega t + \ldots.\, a_n\cos n\omega t + \ldots. \qquad \ldots\ldots(6)$$

We now ignore all harmonic terms assuming that they can be neglected in comparison with the fundamental term. The detailed justification of this step is complicated, and has been considered at some length recently by Russian authors (N. Kryloff and N. Bogoliouboff (1943)). We then obtain on substituting in (4)

$$(i\omega)^m\Delta(i\omega)a_1 + (i\omega)^m\Delta_1(i\omega)A = 0 \qquad \ldots\ldots(7)$$

or

$$Z(i\omega) = -A/a_1, \qquad \ldots\ldots(8)$$

where

$$Z(i\omega) \equiv \Delta(i\omega)/\Delta_1(i\omega)$$

is the output impedance presented to the generator.

If we equate real and imaginary parts of (8) we obtain

$$\mathrm{Im}\,Z(i\omega) = 0 \qquad \ldots\ldots(9)$$

$$\mathrm{Re}\,Z(i\omega) = R = -A/a_1. \qquad \ldots\ldots(10)$$

Equation (9) determines the possible frequencies.

Equation (10) can be shown to be an expression of the conservation of energy. For by (6)

$$\overline{\cos\omega t\,\chi(A\cos\omega t)} = \tfrac{1}{2}a_1,$$

so that (10) may be written in the form

$$-\overline{v\chi(v)} = -\overline{A\cos\omega t\,\chi(A\cos\omega t)} = A^2/2R = \overline{v^2}/R. \qquad \ldots\ldots(11)$$

The left-hand side of (11) is the mean power supplied by the source during a cycle of oscillation, and the right-hand side is the mean power dissipated in the network. It will be noted that to this order of approximation the frequency is independent of the form of the characteristic of the oscillator.

Initial rise of the oscillations

So far we have only been concerned with the resultant amplitude of the steady oscillation. But the procedure of van der Pol can also be employed, in the case of a sinusoidal oscillation, to deduce a differential equation for the initial curve of rise of the oscillations. Instead of the substitution (5) we now use

$$v = V(t)\cos\omega t = \mathrm{Re}\,[v(t)e^{i\omega t}] \qquad \ldots\ldots(12)$$

---

* The minus sign occurs since the current is out from the generator and not in towards it.

and assume that dV/dt can be neglected in comparison with ω, and similarly for higher derivatives. This method will be of great importance in § 3. We now apply it, by way of an illustrative example, to the simple circuit of figure 3.

It is easy to show that in this case

$$\frac{d^2v}{dt^2} + \left[\frac{1}{CR} + \frac{1}{C}\chi'(v)\right]\frac{dv}{dt} + \omega_0^2 v = 0 \qquad \text{......(13)}$$

$$\omega_0^2 = 1/LC.$$

Using the above procedure, and neglecting harmonics as above, we obtain

$$\frac{dV}{dt} + \frac{1}{C}\left[\frac{V}{2R} - \frac{g(V)}{V}\right] = 0, \qquad \text{......(14)}$$

where $g(V) = -\overline{v\chi(v)} = -\overline{V\cos\omega t\,\chi(V\cos\omega t)}$ = power output. The simplest possible non-linear form for $\chi(v)$ is that used by van der Pol, $\chi(v) = i_0 - b_1 v + b_3 v^3$;

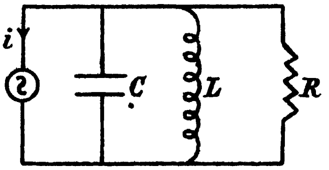


Figure 3.

in this case $g(V)$ equals $b_1 V^2/2 - 3b_2 V^4/8$. We quote the complete solution in this case from van der Pol's paper (1934).

$$V = \sqrt{\frac{4b_1'}{b_3'[1 + e^{-b_1'(t-t_0)}]}}, \qquad b_1' = \frac{b_1}{C} - \frac{1}{CR}, \qquad \text{......(15)}$$

$$b_3' = 3b_3/C.$$

Output circuits

In nearly all practical transmitters, the tuned circuit of the oscillator is coupled to the aerial, which radiates the power, by means of a reactive network. In the case of triode oscillators, the coupling is usually purely electromagnetic; in the case of a magnetron it usually consists of the coupling loop and the waveguide output system.

For any reactive network the solution of (1) at any given frequency ω, can be put in the form

$$Z = z_{11} + \sum_{i=2}^{n} \frac{z_{1i}\Delta_i(i\omega)}{\Delta_1(i\omega)}, \quad \text{where} \quad \Delta_i = \partial\Delta/\partial z_{1i}.$$

All the z_{ij} are pure reactances except $z_{nn} = z$ which is the impedance presented by the aerial feeders. This may be reduced to

$$Z = z_{11} + A(1 + iBz)/(z + iC), \qquad \text{......(16)}$$

where A, B and C are purely real. An exactly analogous procedure is possible for admittances.

As an illustrative example in the case of purely electromagnetic coupling, if M is the mutual inductance and N the self-inductance of the coupling loop

$$Z = z_{11} + \omega^2 M^2/(z + i\omega N), \qquad \text{......(17)}$$

so that $A = \omega^2 M^2$, $B = 0$, $C = \omega N$.

In general, A, B and C are functions of ω, but in most practical cases, such as the one just quoted, they are slowly varying functions of ω, so that if ω varies by only a small fraction of itself they may be regarded as constants.

We now consider in detail the circuit of figure 2.

As our equivalent circuit has been taken with elements in parallel it is convenient to use admittances. We assume that the impedance at the aerial is a resistance $r=1/g$. Let Y_0 be the characteristic admittance of the feeder and l its length. Then admittance across MN presented by the feeder is given by

$$y=Y_0\frac{g\cos\theta+iY_0\sin\theta}{Y_0\cos\theta+ig\sin\theta}=Y_0\frac{\alpha\cos\theta+i\sin\theta}{\cos\theta+i\alpha\sin\theta}, \qquad \ldots\ldots(18)$$

where $\theta=\omega l/c=2\pi l/\lambda$, $\alpha=g/Y_0=$ standing wave ratio.

Taking real and imaginary parts and writing $y=G+iS$ we find

$$G=\frac{1}{R}=\frac{2\alpha Y_0}{1-\alpha^2}\frac{1}{\xi+\cos 2\theta}, \qquad \ldots\ldots(19)$$

$$S=Y_0\frac{\sin 2\theta}{\xi+\cos 2\theta}, \qquad \ldots\ldots(20)$$

where $\xi=(1+\alpha^2)(1-\alpha^2)$.

If we ignore the coupling network completely (i.e. connect straight to the feeder) the frequency would satisfy the relation

$$\omega C-1/\omega L+S=0$$

using (9), or

$$\Delta\omega=-\frac{Y_0}{2C}\frac{\sin 2\theta}{\xi+\cos 2\theta}=\frac{Y_0}{2C}\frac{\sin 2\phi}{\xi-\cos 2\phi}, \qquad \ldots\ldots(21)$$

where

$$\Delta\omega=\omega-\omega_0, \quad \phi=\theta+\pi/2.$$

The power output would be given by equation (11) with

$$\frac{1}{R}=\frac{2\alpha Y_0}{1-\alpha^2}\frac{1}{\xi-\cos 2\phi}. \qquad \ldots\ldots(22)$$

It will be shown a little later that so long as the reactances in the coupling network can be considered constant over the small band of frequencies covered, which is nearly always the case in practice, the coupling network produces no change in the shape of the curves given by (21) and (22), but merelyshifts them or magnifies them. Hence the functions

$$\sin 2\phi/(\xi-\cos 2\phi) \qquad \ldots\ldots(23)$$

and

$$1/(\xi-\cos 2\phi) \qquad \ldots\ldots(24)$$

may be thought of as basic functions in connection with the frequency and power output. They are drawn as functions of ϕ, for various values of ξ in figures 4 and 5.

We also enumerate for convenience several properties of the curves (23) and (24). The maximum of (24) occurs when $\phi=0$ and then $(23)=0$. Similarly the minimum of (24) occurs when $\phi=\pi/2$ and again $(23)=0$. The maximum and minimum of (23) occur when $\cos 2\phi=1/\xi$, and have values

$$\pm\frac{1}{\sqrt{\xi^2-1}}=\pm\frac{1-\alpha^2}{2\alpha}. \qquad \ldots\ldots(25)$$

The slope at the origin of (23) is given by

$$m_0 = \frac{2}{|\xi|-1}. \qquad \ldots\ldots(26)$$

We now deal with a general reactive coupling network and prove the statement made previously. As was shown above, the admittance across JK facing toward the aerial can be put in the form

$$Y = A\frac{1+i\beta y/Y_0}{y+i\gamma Y_0} \qquad \ldots\ldots(27)$$

where A, β, γ are real and assumed constant. Substituting the value of Y from (18) and writing $Y = G+iS$ we obtain after a little simplification

$$G = \frac{A}{Y_0}\frac{2\alpha}{1-\alpha^2}\frac{1+\beta\gamma}{(1+\gamma^2)\xi+2\gamma\sin 2\theta-(1-\gamma^2)\cos 2\theta} \qquad \ldots\ldots(28)$$

and

$$S = \frac{A}{Y_0}\frac{(\beta-\gamma)\xi-(\beta+\gamma)\cos 2\theta-(1-\beta\gamma)\sin 2\theta}{(1+\gamma^2)\xi+2\gamma\sin 2\theta-(1-\gamma^2)\cos 2\theta}. \qquad \ldots\ldots(29)$$

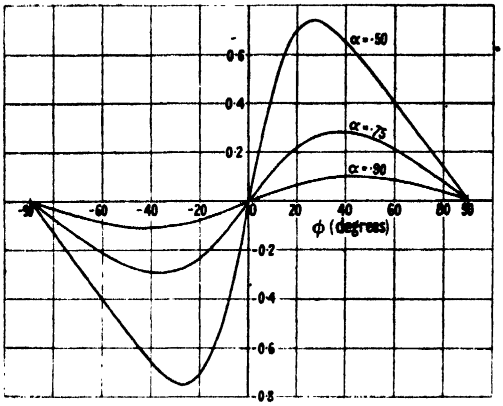


Figure 4.

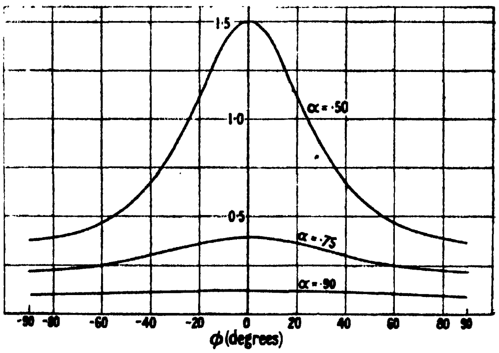


Figure 5.

Write $\gamma = \tan\lambda$; then

$$(1+\gamma^2)\xi+2\gamma\sin 2\theta-(1-\gamma^2)\cos 2\theta = (1+\gamma^2)[\xi-\cos 2(\theta+\lambda)],$$

and it is now easy to show that

$$G = \frac{A}{Y_0}\frac{1+\beta\gamma}{1+\gamma^2}\frac{2\alpha}{1-\alpha^2}\frac{1}{\xi-\cos 2(\theta+\lambda)} \qquad \ldots\ldots(30)$$

and

$$S = \frac{A}{Y_0}\left[\frac{\beta-\gamma}{1+\gamma^2}-\frac{1+\beta\gamma}{1+\gamma^2}\frac{\sin 2(\theta+\lambda)}{\xi-\cos 2(\theta+\lambda)}\right]. \qquad \ldots\ldots(31)$$

The frequency and power output are given by equations similar to (21) and (22).

If we plot the right-hand side and left-hand side of (21) as functions of ω, the intersections of the resulting curves determine the possible frequencies. It is more convenient to plot them as functions of θ. Then the right-hand side is independent of the length of guide, and the left-hand side $= c(\theta-\theta_0)/l$ is a straight line whose slope is inversely proportional to l. For sufficiently small l the curves intersect in only one point. When l increases beyond the critical value given by

$$c/l = Y_0 m_0/2C \qquad \ldots\ldots(32)$$

three frequencies become possible for certain values of the angle θ_0, or for certain phase lengths of the feeder. As l increases still further five frequencies become possible, and so on; this question is discussed in greater detail in a companion experimental paper by B. W. Lythall (1946).

Analysis of experimental data

If the expressions (30) and (31) are examined closely, it will be seen that a considerable similarity exists between the coefficients of the basic frequency and power functions. In fact, if we are given the curve of S as a function of θ, and α is known, we can deduce the value of G at any point of the curve. In practice we do not know S but $\Delta\omega = -S/2C$; hence we can deduce $G/C = 1/CR$ at any point of the curve. Thus using (25) we deduce from (31) that

$$p = (\Delta\omega)_{\max} - (\Delta\omega)_{\min} = \frac{1-\alpha^2}{2C}\frac{A}{Y_0}\frac{1+\beta\gamma}{1+\gamma^2} \qquad \ldots\ldots(33)$$

and hence from (30)

$$\frac{1}{CR} = p\left(\frac{2\alpha}{1-\alpha^2}\right)^2\frac{1}{\xi - \cos 2(\theta+\lambda)}. \qquad \ldots\ldots(34)$$

The above analysis applies satisfactorily to triode oscillators, the function $\chi(v)$ being related to the characteristics of the valve. The mechanism which produces oscillations in the magnetron is different in character, and a detailed analysis is a complicated problem in electromagnetic theory. However it is interesting to apply the present analysis empirically, and it will be seen to yield satisfactory results.

In figure 6 the experimental curves of B. W. Lythall (1946) are reproduced giving frequency and power as a function of θ for various values of α for a magnetron. Theoretical frequency curves are also drawn and it will be seen that they fit the experimental ones quite well. From these curves, using (34), we can plot the power output P as a function of $1/CR$; and similarly if V is the R.F. peak voltage we can plot $V\sqrt{C}$ as a function of $1/CR$. This is done in figure 7 for the given data. Hence we can plot P as a function of $V\sqrt{C}$, and determine, except for a constant, the function $g(V)$ of equation (14). This is done in figure 8, and can be used in the problem of the dynamic case.

The functions in figure 7 and figure 8 can be fitted over a good deal of their range by a very simple empirical formula; this has been previously referred to by American authors. It can be most easily demonstrated by plotting V, the R.F. peak voltage, against I, the R.F. peak current. Actually we plot $\sqrt{2PCR} = V\sqrt{C}$ against $\sqrt{2P/CR} = I/\sqrt{C}$ in figure 9, and will it be seen that the resulting points approximate quite closely, over a good deal of their range, to a straight line. Hence we may write

$$V = V_0 - R_0 I \qquad \ldots\ldots(35)$$

where, from figure 9,

$$\omega CR_0 = 44 \quad V_0\sqrt{C} = 8 \text{ ($\Delta\omega$ being measured in Mc./sec.)} \qquad \ldots\ldots(36)$$

and

$$g(V) = (V/2R_0)(V_0 - V) \qquad \ldots\ldots(37)$$

over the region where (35) is satisfied. This formula should not be extrapolated beyond the region where it is checked by experiment.

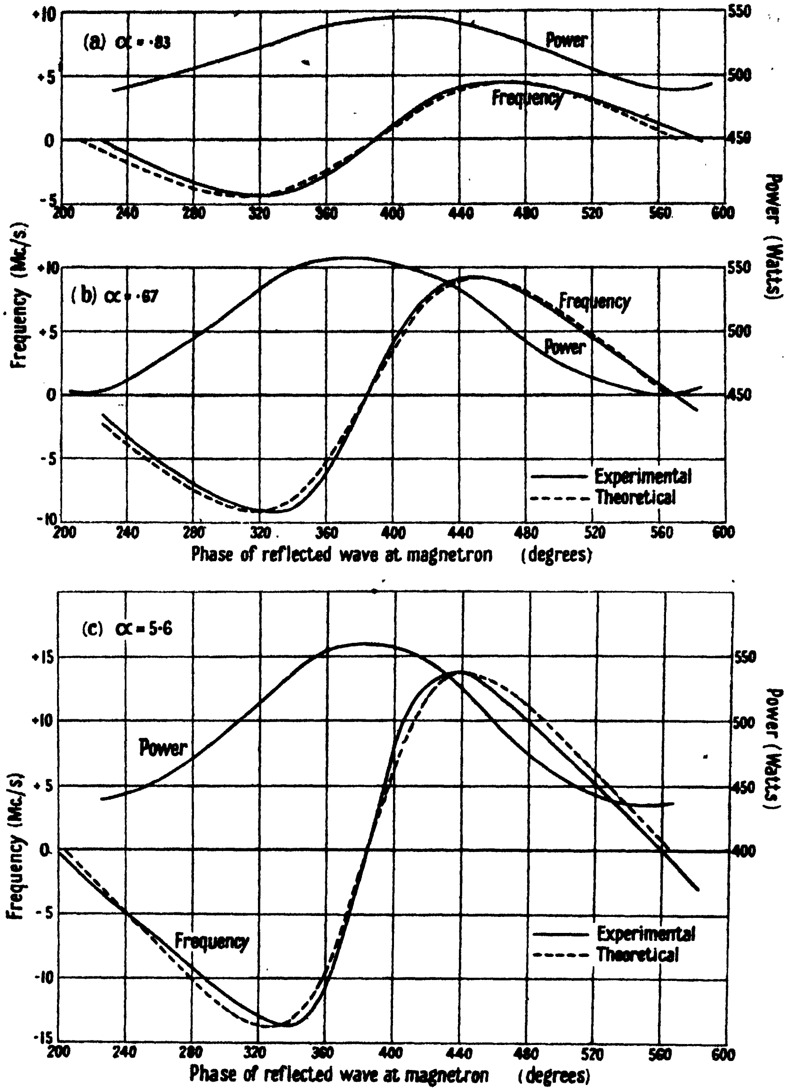


Figure 6.

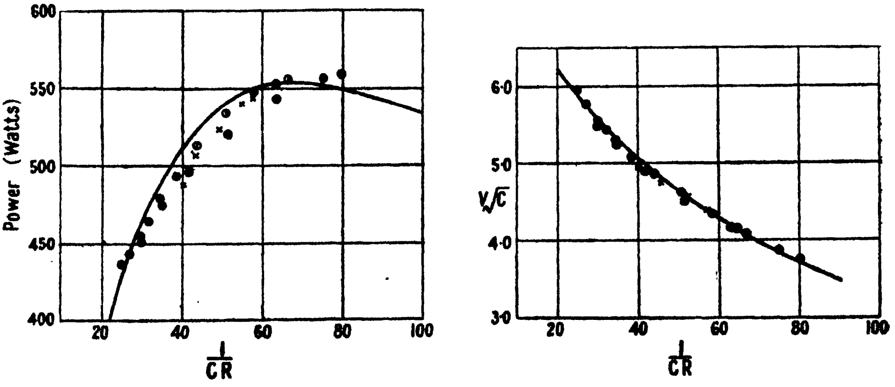


Figure 7.

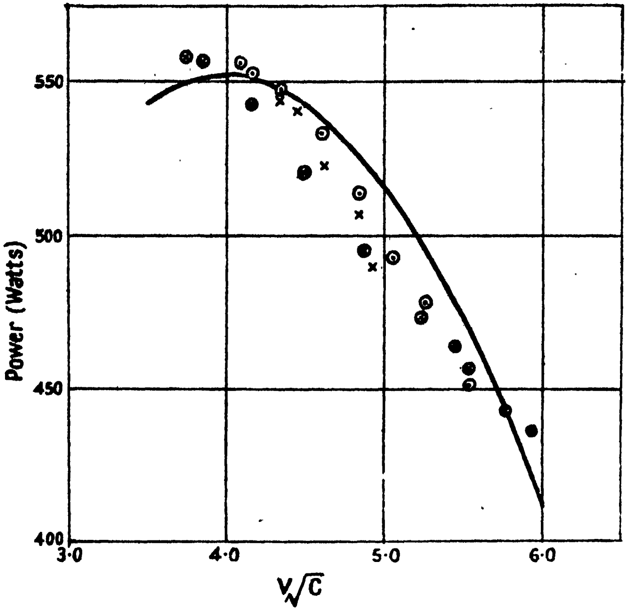


Figure 8.

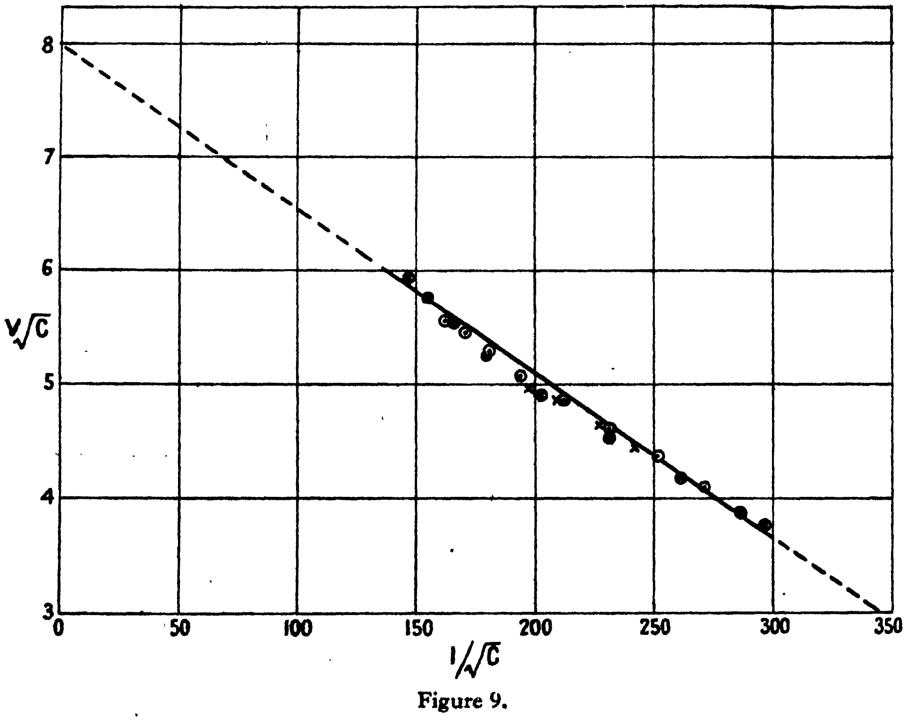


Figure 9.

The Q of an oscillator

The problem of an oscillation generator is non-linear, and is much more complicated than that of a linear network, to which we can apply an unambiguous definition of a Q. A good deal of confusion can thus be caused by using the term loosely without further qualification as to what is meant. There are certainly three different phenomena to which a Q is applicable, all depending on the load.

(1) The Q of the oscillating system itself without the oscillator; if the oscillator were switched off suddenly, the rate of decay of the power would enable this Q to be determined. In terms of the simple equivalent circuit of figure 3, $Q = 1/\omega CR$. Typical values for the experimental results on the magnetron can be read off from the abscissa of figure 7 (assuming the guide to be sufficiently short for the propagation time to be negligible in comparison with the time of decay). The units are chosen so that we obtain Q by dividing 3000 by the appropriate value of $1/CR$; hence for the experimental points taken Q varies between 37 and 120.

(2) If an oscillation is taking place at a certain stable amplitude A and is disturbed at time t_0 to amplitude $A+x$ then at time t the amplitude will be $A+xe^{-\delta(t-t_0)}$; this gives rise to another Q. For the magnetron, this Q may be estimated by use of the empirical results of equation (35). Over the experimental range covered it varies between 20 and 32.

(3) The initial rate of rise of the oscillations also determines a Q.

§ 3. THE DYNAMIC CASE

We now propose to deduce rigorously the complete equations for the circuit of figure 2. We ignore the coupling network and assume that the output circuit is directly coupled to the feeder; this introduces some simplification, and the considerations of the the static case seem to indicate that the assumption involves little loss of generality. We then show that the steady solutions of the equations are exactly the same as those deduced in the static case; and we derive approximations which are useful in the case when the amplitude modulation is small and the frequency modulation is of prime importance.

Let the voltages and currents be as labelled in figure 10.

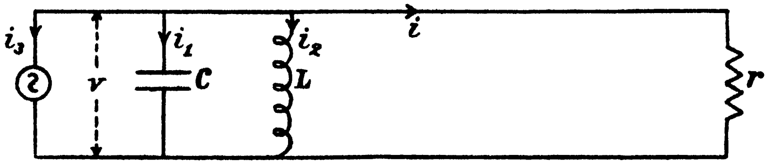


Figure 10.

Then
$$v = L\frac{di_2}{dt} = \frac{1}{C}\int i_1 dt$$

and
$$i_3 = -(i+i_1+i_2) = \chi(v).$$

Hence
$$\frac{d^2v}{dt^2} + \frac{1}{C}\frac{d}{dt}[\chi(v)] + \omega_0{}^2 v = -\frac{1}{C}\frac{di}{dt}. \qquad \ldots\ldots(38)$$

From transmission line theory we have

$$\left.\begin{aligned} v &= f(t) - \rho f(t-\tau), \\ Z_0 i &= f(t) + \rho f(t-\tau), \end{aligned}\right\} \qquad \ldots\ldots(39)$$

where $\alpha = \dfrac{Z_0}{r} = \dfrac{g}{Y_0}$ as above, $\rho = \dfrac{1-\alpha}{1+\alpha}$,

τ = propagation time forwards and backwards along the line. We deal with equation (38) by the method described in § 2.

Write

$$\left.\begin{aligned} v &= \mathrm{Re}\,(Ve^{i\omega_0 t}) = v_1 \cos\omega_0 t - v_2 \sin\omega_0 t \\ &= |V| \cos(\omega_0 t + \eta), \\ i &= \mathrm{Re}\,(Ie^{i\omega_0 t}), \\ f(t) &= \mathrm{Re}\,[F(t)e^{i\omega_0 t}], \end{aligned}\right\} \qquad \ldots\ldots(40)$$

where $|V|$ and η are slowly varying in comparison with ω_0.

We expand $\chi(v) = \chi[|V| \cos(\omega_0 t + \eta)]$

in terms of cosine harmonics of $(\omega_0 t + \eta)$ and retain only the first harmonic. Thus we write $\chi(v)$ as approximately equal to

$$-\frac{2g(|V|)}{|V|} \cos(\omega_0 t + \eta),$$

where $g(|V|)$ is defined by (14). Hence

$$\frac{d}{dt}[\chi(v)] = -2\frac{d}{dt}\left[\frac{g(|V|)}{|V|}\right] \cos(\omega_0 t + \eta) + (\omega_0 + \eta')\frac{2g(|V|)}{|V|} \sin(\omega_0 t + \eta),$$

and neglecting terms of order $\dfrac{\eta'}{\omega_0}$, $\dfrac{1}{\omega_0}\dfrac{d|V|}{dt}$ we may write this as

$$\mathrm{Re}\,\frac{2i\omega_0 g(|V|)}{|V|^2} Ve^{i\omega_0 t}.$$

Proceeding similarly with the other terms of (38) we obtain, after a little simplification,

$$\frac{dV}{dt} - \frac{g(|V|)}{C|V|^2} V = -\frac{1}{2C} I. \qquad \ldots\ldots(41)$$

Applying the substitution (40) to (39), we have

$$\left.\begin{aligned} V &= F(t) - \rho e^{-i\omega_0 \tau} F(t-\tau), \\ Z_0 I &= F(t) + \rho e^{-i\omega_0 \tau} F(t-\tau). \end{aligned}\right\} \qquad \ldots\ldots(42)$$

Hence $Z_0 I = F(t) + \rho e^{-i\omega_0 \tau} F(t-\tau)$,

and (41) becomes

$$\frac{dV}{dt} + \left[\frac{1}{2CZ_0} - \frac{g(|V|)}{C|V|^2}\right] V = -\frac{\rho}{CZ_0} e^{-i\omega_0 \tau} F(t-\tau). \qquad \ldots\ldots(43)$$

If we take real and imaginary parts of (43) we have a pair of equations which in conjunction with the first equation of (42) completely determine the form of the oscillation generated. Actually although this form of the equation is convenient for integration on a differential analyser, it is difficult to see the physical significance of the various terms. For the purpose of physical approximation it is more useful to transform the equation into a pair involving amplitude and phase explicitly. Taking the conjugate complex of (43),

$$\frac{d\bar{V}}{dt} + \left[\frac{1}{2CZ_0} - \frac{g(|V|)}{C|V|^2}\right] \bar{V} = -\frac{\rho}{CZ_0} e^{-i\omega_0 \tau} \bar{F}(t-\tau). \qquad \ldots\ldots(44)$$

Multiplying (43) by $\overline{V}$, (44) by V and adding,

$$\frac{d}{dt}(|V|^2)+2(|V|^2)\left[\frac{1}{2CZ_0}-\frac{g(|V|)}{C|V|^2}\right]=-\frac{2\rho}{CZ_0}[v_1k_1+v_2k_2].$$

Similarly subtracting,

$$2i\frac{d}{dt}\tan^{-1}\frac{v_2}{v_1}=-\frac{2i\rho}{CZ_0}\frac{v_1k_2-v_2k_1}{|V|^2}.$$

where

$$e^{-i\omega_0\tau}F(t-\tau)=k_1+ik_2=K(t). \qquad \ldots\ldots(45)$$

These equations can be written as

$$\left.\begin{aligned}\frac{d|V|}{dt}+\left[\frac{|V|}{2CZ_0}-\frac{g(|V|)}{C|V|}\right]&=-\frac{\rho}{CZ_0}[k_1\cos\eta+k_2\sin\eta],\\ \frac{d\eta}{dt}&=-\frac{\rho}{CZ_0|V|}[k_2\cos\eta-k_1\sin\eta].\end{aligned}\right\} \qquad \ldots\ldots(46)$$

As a check on the correctness of equations (46) we find the values of amplitude and frequency for which steady solutions are possible, and compare them with the results obtained in the discussion of the static case. The conditions for a steady solution are

$$d\eta/dt=\Delta\omega, \quad d|V|/dt=0.$$

Hence from (42)

$$F=\frac{V}{1-\rho e^{-i\omega\tau}}, \quad \omega=\omega_0+\Delta\omega.$$

Substituting in (46),

$$\Delta\omega=-\frac{\rho}{CZ_0}\,\mathrm{Im}\left(\frac{e^{-i\omega\tau}}{1-\rho e^{-i\omega\tau}}\right)=-\frac{1}{2CZ_0}\frac{\sin\omega\tau}{\xi+\cos\omega\tau}, \qquad \ldots\ldots(47)$$

where

$$\xi=-(1+\rho^2)/2\rho=(1+\alpha^2)/(1-\alpha^2)$$

and

$$\frac{|V|}{2CZ_0}-\frac{g(|V|)}{C(|V|)}=-\frac{\rho|V|}{CZ_0}\,\mathrm{Re}\left(\frac{e^{-i\omega\tau}}{1-\rho e^{-i\omega\tau}}\right)$$

or

$$\frac{g(|V|)}{|V|^2}=\frac{1}{2Z_0}\frac{2\alpha}{1-\alpha^2}\frac{1}{\xi+\cos\omega\tau}. \qquad \ldots\ldots(48)$$

Equations (47) and (48) will be found to be analogous to (21) and (22).

Approximations to the equations

Equations (46) are cumbersome, even for numerical integration, because of the complex difference relations. From the physical point of view, the best method of solution would appear to be one of successive approximations. Often only a qualitative picture is required for any given conditions, and then the first approximation is usually sufficient. Also in practice, the amplitude modulation is often small, so that only the frequency modulation need be considered to the first order.

From the first equation of (42) we deduce that

$$F(t)=V(t)+\rho e^{-i\omega_0\tau}V(t-\tau)+\rho^2e^{-2i\omega_0\tau}V(t-2\tau)+\ldots.$$

the series continuing until $V(t-n\tau)=0$.

Hence

$$\rho K(t)=\rho e^{-i\omega_0\tau}V(t-\tau)+\rho^2e^{-2i\omega_0\tau}V(t-2\tau)+\ldots.$$

and

$$\frac{\rho K(t)}{V}=\rho e^{-i[\omega_0\tau+\eta(t)-\eta(t-\tau)]}\left|\frac{V(t-\tau)}{V(t)}\right|+\rho^2e^{\;i[2\omega_0\tau+\eta(t)-\eta(t-2\tau)]}\left|\frac{V(t-2\tau)}{V(t)}\right|$$

$$+\ldots.\quad+\rho^re^{-i[r\omega_0\tau+\eta(t)-\eta(t-r\tau)]}\left|\frac{V(t-r\tau)}{V(t)}\right|+\ldots. \qquad \ldots\ldots(49)$$

We now approximate by assuming in (49) that all terms such as $\left|\frac{V(t-r\tau)}{V(t)}\right|$ can be put equal to unity. We put the resulting function equal to $H(t)$, so that

$$H(t)=h_1(t)+ih_2(t)=\rho e^{-i[\omega_0\tau+\eta(t)-\eta(t-\tau)]}+\rho^2e^{-i[2\omega_0\tau+\eta(t)-\eta(t-2\tau)]}+\ldots. \qquad \ldots\ldots(50)$$

With this assumption the second equation of (46) is independent of the first and can be written

$$\frac{d\eta}{dt}=-\frac{1}{CZ_0}h_2(t). \qquad \ldots\ldots(51)$$

Also it is easy to see that

$$\left.\begin{aligned}H(t-\tau)&=\frac{H(t)}{\rho e^{-i\psi}}-1,\\ \text{where}\qquad \psi&=\omega_0\tau+\eta(t)-\eta(t-\tau).\end{aligned}\right\} \qquad \ldots\ldots(52)$$

Splitting (52) into its real and imaginary parts,

$$\left.\begin{aligned}h_1(t)&=\rho\cos\psi[1+h_1(t-\tau)]+\rho\sin\psi h_2(t-\tau),\\ h_2(t)&=-\rho\sin\psi[1+h_1(t-\tau)]+\rho\cos\psi h_2(t-\tau).\end{aligned}\right\} \qquad \ldots\ldots(53)$$

Equation (51) in conjunction with (53) is not difficult to deal with numerically. If a solution has been determined and a better approximation is desired it should be possible to achieve this by an iterative method applied to (46). The solution can be substituted in the first equation of (46) to obtain a first approximation to $|V|$. This can be used then in the second equation to obtain a second approximation to η and so on. It seems reasonable to assume that this sequence of approximations converges fairly rapidly.

The condition for a steady solution is that $h(t)=h(t-\tau)$, and then from (51) and (52)

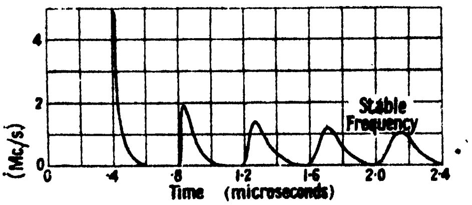
$$\frac{d\eta}{dt}=-\frac{1}{CZ_0}\operatorname{Im}\frac{\rho}{e^{i\psi}-\rho}=-\frac{1}{2CZ_0}\frac{\sin\psi}{\xi+\cos\psi}. \qquad \ldots\ldots(54)$$

This is, in fact, the condition (21) derived for the static case, and is only strictly valid when $d\eta/dt$ and ψ are constant. However, it has been suggested empirically by Mr. C. L. Ratsey that if we use this equation in the general case we should get quite a useful qualitative idea of the resulting frequency. Numerical solutions of equations (52) and (54) for a typical case, which were integrated by Dr. E. T. Goodwin, did indicate qualitative agreement. Equation (54) also has the advantage that it can be put on a small differential analyser, and solutions for many different values of α, $1/CZ_0$, τ, $\omega_0\tau$ can then be run off. This was done on the small differential analyser at Cambridge for values of

$$\alpha = 0{\cdot}5,\ 0{\cdot}68,\ 0{\cdot}95; \qquad \tau = 0{\cdot}1,\ 0{\cdot}25,\ 0{\cdot}4\ \mu\text{sec.};$$
$$1/CZ_0 = 240; \qquad \omega_0\tau = 170^\circ,\ 60^\circ,\ 30^\circ.$$

The functions η and ψ were integrated directly and then $d\eta/dt$, or instantaneous frequency, can be determined from (54). Three typical solutions are reproduced in figures 11, 12, 13. As is expected solutions near $\omega_0\tau = 0^\circ$ are stable and settle down quickly, whereas those in the neighbourhood of $\omega_0\tau = 180^\circ$, where variations in the initial conditions of rise can cause instability, take a long time to attain the stable frequency.

Figure 11



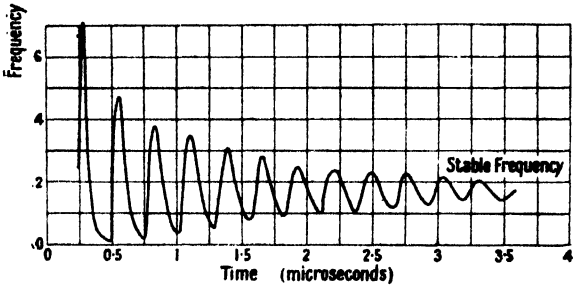


Figure 12.

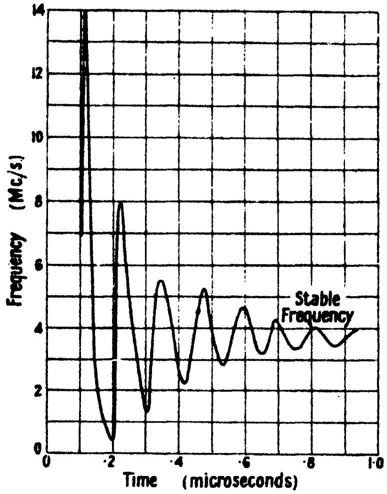


Figure 13.

The experimental results in §2 did not yield $g(|V|)$ directly, but the power output as a function of $|V|\sqrt{C}$, say $g_1(|V|\sqrt{C})$; if we wish to use these results we can still deduce relative amplitudes by the transformation,

$$V_1 = V\sqrt{C}, \quad F_1 = F\sqrt{C}.$$

We then obtain on substituting in (43)

$$\frac{dV_1}{dt} + \left[\frac{1}{2CZ_0} - \frac{g_1(|V_1|)}{|V_1|^2}\right] V_1 = -\frac{\rho}{CZ_0} e^{-i\omega_0\tau} F_1(t-\tau). \qquad \ldots\ldots(55)$$

If we are using the simple formula (37) for $g(V)$ it becomes convenient to express voltages in terms of V_0 as unit. Writing $V/V_0 = V_2$, $F/V_0 = F_2$ (43) assumes the form

$$\frac{dV_2}{dt} + \left[\frac{1}{2CZ_0} - \frac{1}{2CR_0}\left(\frac{1}{|V_2|} - 1\right)\right] V_2 = -\frac{\rho}{CZ_0} e^{-i\omega_0\tau} F_2(t-\tau).$$
$$\ldots\ldots(56)$$

The instability in the neighbourhood of $\omega_0\tau = 0$ and the probability of either solution for any particular value of $\omega_0\tau$ has not been discussed theoretically. It must involve a detailed consideration of the initial conditions of rise of the oscillator, and is a problem of considerable difficulty.

ACKNOWLEDGMENTS

The work described was carried out in the Admiralty Signal Establishment, and the paper is published with the approval of the Board of Admiralty. The author is indebted to the Captain Superintendent for granting facilities for the preparation of the paper. He also wishes to express his thanks to Dr. W. T. Davies for drawing his attention to this problem and for many useful discussions concerning it; to Dr. E. T. Goodwin for his advice and help in connection with the approximations and numerical integrations; to Mr. J. Crank and the staff of the Mathematics Laboratory at Cambridge for their work on the differential analyser; and to Mr. B. W. Lythall for providing experimental data.

REFERENCES

Appleton and Greaves, 1923. *Phil. Mag.*, **45**, 401.
Kryloff and Bogoliouboff (free translation by S. Lefschetz), 1943. *Introduction to Non-Linear Mechanics* (Princeton University Press).
Lythall, 1946. *J. Instn. Elect. Engrs.*, **93**, IIIA, 1081.
van der Pol, 1922. *Phil. Mag.*, **43**, 700.
van der Pol, 1934. *Proc. Inst. Radio Engrs.*, **22**, 1051.

ON THE FORMATION OF HEAVY ELEMENTS IN STARS

By F. HOYLE,
Cambridge

Read 20 *December* 1946 ; *MS. received* 13 *March* 1947

§ 1. INTRODUCTION

Astrophysical data indicate that hydrogen and helium together are about ten thousand times as abundant as all other elements combined. This statement is based on Russell's general discusssion of the opacity of stellar atmospheres (1933), Strömgren's detailed analysis of the solar atmosphere (1940) and Dunham's estimate for the composition of the interstellar gas (1939). The helium is regarded as arising through synthesis from hydrogen, either by the carbon-nitrogen cycle or by deutron formation (Bethe and Critchfield, 1938; Bethe, 1939). On this basis hydrogen must have been by far the most abundant element during the early history of the universe (it is generally believed that in spite of the subsequent formation of helium, hydrogen is still the most abundant element). It is usually regarded as a natural extrapolation to pass from this conclusion (which is forced on us by observation) to the hypothesis that hydrogen was the only element initially present in the universe. It is then necessary to explain how the heavy elements have been synthesized from hydrogen. The present paper is concerned with a recent attempt to solve this problem (Hoyle, 1947).

It will be useful to divide the discussion into two parts. First we consider the physical conditions necessary for the production of the heavy elements, and second the place in the universe where such conditions are realized will be described. The remainder of the present section, together with § 2, will be devoted to the first part, and we shall return to the astronomical considerations in § 3.

There are two different ways of approaching the physical part of the problem. We could attempt to synthesize a particular element by specifying in detail a chain of nuclear reactions that start from hydrogen and end with the element in question. Such a method is at first sight attractive, because it can be directly related to processes observed in the laboratory. But a non-statistical method breaks down in attempting to proceed in a reverse direction along the radioactive series beyond such very short-lived nuclei as RaC′ and ThC′. This difficulty does not arise in a statistical treatment, and for this reason alone it would be preferable to adopt the method of statistical mechanics (in which it is not necessary to specify individual processes in detail).

§ 2. THE PHYSICAL PART OF THE PROBLEM

It is well known in the theory of dissociating gases that if for the *n* substances $A_1, A_2, \ldots A_n$ it is possible to find a chain of reactions that connects A_r with A_s (for all pairs *r* and *s*) then the relative abundances of the substances can be determined entirely from statistical mechanics. The statistical treatment assumes that the system is given sufficient time to reach equilibrium. This proviso is very important in the present problem, because it restricts the search for a suitable place where the synthesis of the elements may occur to a particular class of exceptionally dense, hot stars. The astronomical considerations of § 3, taken together with quantitative calculations of the speed of nuclear reactions, show that statistical equilibrium over the whole periodic table requires temperatures greater than about 4.10^9 °c. (which may be compared with a central temperature in the Sun of about 2.10^7 °c.). We shall confine attention to temperatures exceeding this value.

The equations governing the statistical equilibrium between nuclei are closely similar to the equations for dissociating gases (Sterne, 1933; Fowler, 1936). When the density ρ and the temperature T are given, the equations determine the abundances of the elements. Thus for each pair of values of ρ, T there is a unique composition for material under statistical equilibrium. If we regard $\log_{10}\rho$ (ρ in gm. per cm^3) and T (in units of 10^9 °c.) as rectangular Cartesian coordinates, there will be a definite composition for material in statistical equilibrium at a given point in the ρ, T plane.

The first general result given by the statistical equations is that the ρ, T plane can be divided into two parts by the curve AEFB of figure 1, which is such that to the right of the curve the material is almost entirely composed of helium, while to the left of the curve the material is largely composed of elements with atomic weights greater than 50.

The composition differs appreciably from one part of the heavy element zone to another. As an example the following table gives the composition at D, which is the point $\rho = 10^7$ gr. per cm^3, $T = 4.10^9$ °c.:—

Table 1

Element	^{4}He	^{16}O	^{28}Si	^{56}Fe	^{63}Cu	^{82}Kr	^{118}Sn	^{208}Pb
Logarithm to base 10 of number of nuclei per cm^3	27·2	20·2	23·7	28	26·7	18·8	−3·3	−222

The negligible abundances of elements in the upper half of the periodic table is a striking feature of these values. This result is characteristic of points near D. There is no hope of obtaining the heaviest elements from the region of the ρ, T plane in the neighbourhood of D. For this purpose it is necessary to discuss the composition of material at much higher densities.

Increasing density produces the following important effects:—

(1) A rapidly increasing density of free neutrons.

(2) A rapidly increasing ratio of free neutron density to free proton density.

The increasing ratio of free neutrons to free protons has the effect of increasing the abundances of nuclei containing appreciably more neutrons than protons. That is, the abundances of elements at the upper end of the periodic table are increased at the expense of elements in the lower half of the table. This effect is illustrated by the following values:—

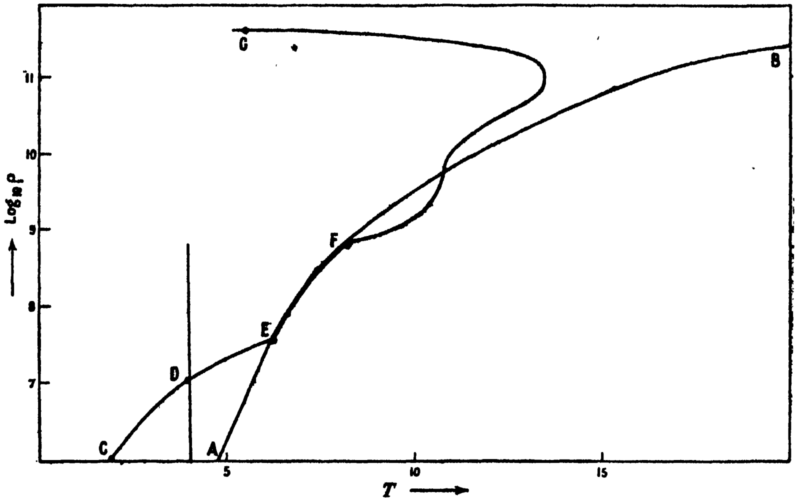


Figure 1. T in units of 10^9 °c. ρ in gr. per cm^3

Table 2

$\log_{10} \rho'$	11·3	11·5	11·6	11·8
T	9·8	11·6	12·5	14·5
Atomic weight of the most abundant element	130	160	180	240

The quantity ρ' is the difference between the total density ρ of the material and the density of free neutrons (which is comparable with, but less than, ρ at these very high densities). Although it is outside the scope of the present paper to discuss the calculations leading to the values in table 2, it may be noted that the energy of relativistically degenerate electrons is the main factor that controls the properties of matter at very high densities.

The above discussion may be summarized by noting that in the neighbourhood of the point D of figure 1 the composition of material is almost entirely restricted to the lower half of the periodic table, whereas in the neighbourhood of G the composition is mainly confined to the upper half of the periodic table.

§ 3. THE ASTRONOMICAL PART OF THE PROBLEM

The discussion of § 2 narrows considerably the search for places where the synthesis of heavy elements can occur. Normal stars, in which energy production through the conversion of hydrogen to helium balances the loss of radiation at the surface, are known from the theory of stellar constitution to possess central temperatures that are lower than the present requirements by a factor of about a hundred.

The Sun possesses hydrogen sufficient to maintain the present normal state for a time of about 10^{11} years, which is considerably greater than the usual estimates for the age of the universe. Accordingly there is no prospect of a departure from the normal state in small stars such as the Sun. The situation is very different, however, for massive stars. The energy production necessary to balance the loss of radiation from the surface is so large in a star of twenty solar masses, for example, that the supply of hydrogen becomes exhausted in about 10^8 years (unless further supplies of hydrogen are added by a rapid sweeping up of the interstellar gas). If the hydrogen should become exhausted, energy production by the carbon-nitrogen cycle, and by any other process involving proton reactions, ceases. On the other hand the loss of energy by radiation at the surface *must* continue as before. This loss of energy then leads to a slow contraction of the star. Now it is well known in the theory of the constitution of the stars that the central temperature of a *slowly* collapsing star is approximately inversely proportional to the radius. Thus the central temperature rises as the radius decreases, and it is clear that if sufficient contraction occurs in massive stars that have exhausted their internal supply of hydrogen, then temperatures of order $5 \,.\, 10^9$ °C. must be attained.

By good fortune it is possible to make a quantitative test of this suggestion. From the work of Baade (1942) and Minkowski (1942) on the supernova of A.D. 1054 we can obtain estimates for the mass and radius of the star before the supernova outburst occurred. These values are about fifteen solar masses, and about one twentieth of the solar radius, respectively. Since the central temperature is proportional to the ratio of mass to radius, its value in this case must be about 300 times that of the Sun, or about $6 \,.\, 10^9$ °C., which is in excellent agreement with our requirements.

We next consider the properties of collapsing stars in further detail. Chandrasakhar has shown that if the mass of a contracting star exceeds about 1·5 solar masses, then the star cannot attain a spherically symmetric equilibrium state, no matter how long contraction continues. This means that contraction would continue in a non-rotating star until the radius became comparable with the gravitational radius (which is only a few kilometres, even for massive stars). On the other hand, in a rotating star, centrifugal forces become important as the contraction continues, and a stage must be reached at which the rotational energy of the star becomes so large that its shape becomes markedly spheroidal. In the equatorial plane further contraction is prevented, but contraction perpendicular to this plane will continue if the star is sufficiently massive until the polar radius is about half the equatorial radius. As Jeans showed, the star now becomes unstable, and a sharp edge develops in the equatorial plane, through which material is ejected. The motion of this material is very difficult to trace by theoretical analysis. It is reasonable, however, to assume that the material behaves in accordance with the observed motions of material ejected by novae and supernovae.

The degree of contraction required to produce instability depends on the amount of angular momentum possessed by the star, and this is known to vary widely from one star to another. Thus if a star possesses a large amount of angular momentum (corresponding to an equatorial rotational velocity in the normal state of several hundred km. per sec.), the degree of contraction required to induce rotational instability is small, and the velocity of ejection of material is small (several hundred km. per sec.). On the other hand, if we take a massive star with the same angular momentum per unit mass as the Sun, for example, the degree of contraction is very large indeed, and the velocity of ejection is correspondingly large (several thousand km. per sec.). So we have a whole range of cases from mild to extremely violent outbursts, according to the angular momentum of the star. It seems possible to fit a variety of observed phenomena such as Wolf-Rayet emission, novae and supernovae in this range.

§ 4. THE SYNTHESIS OF THE HEAVY ELEMENTS

We can now describe the changes that occur in the composition of material in a collapsing star. The curve CDEFG of figure 1 is a representative case worked out from the theory; it has been shown that material near the centre of a collapsing star of five solar masses must evolve along this track. Before reaching D the material is mainly composed of helium. Nuclear reactions become sufficiently rapid for statistical equilibrium to be attained in the neighbourhood of D, and between D and E the helium is converted into a distribution of the form shown in table 1.

This synthesis of heavy elements from helium yields energy, which is then available to balance the loss of radiation at the surface. Accordingly the collapse of the star is temporarily arrested by this energy production, but is resumed when the conversion of the helium is completed. The delay in the contraction may last as long as a million years. This delay is in striking contrast with the behaviour of the star after the point E has been reached. For after passing E the internal pressure is inadequate to provide even approximate support of the star against gravity. An extremely rapid collapse ensues, leading to an evolution from E to G in about a hundred seconds. The reason for this remarkable behaviour is readily understood. For it is clear that the direction of the tangent to the curve must change discontinuously at E, since no energy is available at E to reconvert into helium the heavy elements synthesized between D and E. Such a reconversion requires a supply of energy that can only be forthcoming from a rapid collapse of the star. During this contraction the material follows a section of the curve bounding the helium zone. Thus during the evolution along the section EF of the track, the star is supplying gravitational energy which is absorbed in the reconversion of heavy elements to helium. The reconversion is completed at F, and the material can now enter the helium zone. It might seem that the star can now return to a state of approximate mechanical equilibrium. This is not the case, however, on account of the increasing importance of free neutron production which leads to a further absorption of large quantities of energy. The form of the track between F and G is due to the effect of free neutron production taken together with the non-linear properties of the equations of statistical equilibrium. The sharp decrease of temperature in the neighbourhood of G slows down the rate of the nuclear reactions to such an extent that it is doubtful whether the statistical equations are adequate to describe the form of the track beyond G.

It is tacitly assumed in the above discussion that the evolution of material along CDEFG is not interrupted by the onset of rotational instability. The remarks of the previous section show that the stage at which the evolution is interrupted depends on the angular momentum of the star. If the angular momentum is large enough, instability will occur before material at the centre of the star has evolved beyond E. In this case the synthesis of heavy elements is confined to the lower half of the periodic table. On the other hand, in stars with sufficiently small angular momentum, the material can reach the neighbourhood of G before the onset of instability. The slowing down of the rate of nuclear reactions near G is strongly indicative that in this case the composition of material will remain frozen during the instability process.

It is seen, therefore, that elements in the lower half of the periodic table are provided by material near D, whereas the elements in the upper half of the table require the material to reach the neighbourhood of G before the instability occurs. In this connection it may be noted that the central density of a star is about thirty times the mean density. Consequently there must be a considerable variation in the composition of material in different parts of a rotationally unstable star.

§ 5. GENERAL REMARKS

The theory described above, when taken together with the hypothesis that hydrogen was the only element initially present in the universe requires the first stars to be initially composed of pure hydrogen. It is known from the study of stellar structure that the time required for a massive star to exhaust its supply of hydrogen is substantially independent of the mode of conversion of the hydrogen to helium. Thus about 10^8 years after the formation of the first massive stars, it is to be expected that heavy elements would begin to be distributed in interstellar space. Subsequent stellar condensations are formed from the initial hydrogen together with these heavy elements. It follows, therefore, that pure hydrogen stars condense only in the earliest stages in the evolution of a galaxy. The relation of this question to the origin of the white dwarf stars has been discussed in a recent paper (Hoyle, 1947 b).

A star that becomes rotationally unstable after the central material reaches the point E in the ρ, T plane must undergo an extremely violent outburst, since the unstable state is reached in a catastrophic manner. On the other hand, a star that becomes unstable before the central material arrives at E takes many thousands of years to reach the unstable state. The critical nature of the point E provides a natural explanation of the difference between novae and supernovae. That is, supernovae are collapsed stars that become unstable *after* the point E has been reached, whereas novae are collapsed stars that reach the unstable state *before* the point E is attained.

It remains to show that the present theory is capable of supplying the required quantity of heavy elements. The average rate at which elements are distributed in interstellar space is evidently of the same order as the mean rate at which material is ejected by supernovae. Using the observational estimate of one supernova per galaxy per 500 years we obtain, on the basis of the ejection of ten solar masses per supernova, about $2 \,.\, 10^8$ solar masses distributed as heavy elements in interstellar space in 10^{10} years (this time is in accordance with the usual estimates for the age of the universe). This gives an abundance of heavy elements amounting to about 0·1% by mass of the hydrogen abundance. The remarks made at the outset, concerning the relative abundance of heavy elements and hydrogen, show that this estimate is of the required magnitude.

REFERENCES

Baade, W., 1942. *Astrophys. J.*, **96**, 188.
Bethe, H. A., 1939. *Phys. Rev.*, **55**, 434.
Bethe, H. A., and Critchfield, C. L., 1938. *Phys. Rev.*, **54**, 248.
Dunham, T., Jr., 1939. *Proc. Amer. Phil. Soc.*, **81**, 277.
Fowler, R. H., 1936. *Statistical Mechanics* (Cambridge : The University Press), p. 655.
Hoyle, F., 1947 a. *Mon. Not. R. Astr. Soc.* (in course of publication).
Hoyle, F., 1947 b. *Mon. Not. R. Astr. Soc.* (in course of publication).
Minkowski, R., 1942. *Ap. J.*, **96**, 199.
Sterne, T. E., 1933. *Mon. Not. R. Astr. Soc.*, **93**, 736.
Strömgren, B., 1940. *Festschrift für Elis Strömgren.*
Russell, H. N., 1933. *Astrophys. J.*, **78**, 239.

DISCUSSION

Dr. Martin Johnson. The following difficulties occur to me : All known stars exhibit hydrogen in considerable abundance, and yet Dr. Hoyle's mechanism can only begin when all hydrogen has been exhausted. Heavy elements are also observed in the same stars though their genesis ought to have come via Dr. Hoyle's mechanism. It seems likely that most stars entering on his mechanism of collapse would be instantly re-expanded by their remaining hydrogen before his densities and temperatures are reached.

Prof. R. E. Peierls. One should bear in mind that in non-equilibrium conditions, the formation of heavy elements may be possible at pressures and temperatures at which the amount of such elements in thermal equilibrium would be negligible. To use an analogy, the distribution over the surface of the earth of organic compounds probably does not correspond to the abundances they would have at any temperature or pressure. If, in the course of reactions between light elements, free neutrons are produced, they may lead to a gradual building up of heavier elements.

Most of the discussions of abundance seem to start from the hypothesis that the primary element is hydrogen and in some stage the universe consisted of hydrogen only. This is the simplest, but by no means the only, possible hypothesis, and if we explain the origin of heavy elements by their formation from hydrogen the next question is evidently, where does the hydrogen come from ?

Author's reply. There is no difficulty in understanding how both hydrogen and heavy elements may come to be present in the same star. All that is required is for the star to condense *after* the heavy elements began to be distributed in interstellar space. Moreover, even if a star were initially composed of pure hydrogen it could still acquire heavy elements through the subsequent accretion of interstellar material.

Stars such as V Pupp. and Y Cygn. must exhaust their internal supply of hydrogen in about 10^8 years, unless they are replenished with hydrogen by accretion. Thus in a time that is, astronomically speaking, very short, a number of very bright stars, now observed in a normal state, must evolve along the collapsing sequence that leads to the processes I have described. The situation is even more marked in the case of 29 Can. Maj. and Ao Cass., where the hydrogen supply, if unreplenished, will be exhausted in about 10^7 years.

In using the analogy of organic compounds it must be remembered that the high-grade radiation from the Sun plays an essential part in the building of these compounds. The presence of a corresponding high-energy source would be necessary in a non-statistical approach to the present problem. Even if such a source were forthcoming I think great difficulty would still arise in attempting to pass the barriers at RaC′ and ThC′.

The observed high abundance of hydrogen is strongly suggestive that only hydrogen was initially present in the universe. The question of the origin of the hydrogen may well be of great significance, but this problem must be left over for future discussion.

NUCLEAR MAGNETIC MOMENTS

By R. LATHAM,
Cavendish Laboratory, Cambridge

MS received 20 *March* 1947

ABSTRACT. A regular change in the nuclear magnetic moments of the elements as a function of their charge and spin is described, and a connection is suggested between this and the excess of the number of neutrons over the number of protons in the nucleus.

§ 1. INTRODUCTION

The purpose of this note is to point out an empirical regularity of behaviour in the nuclear magnetic moments of the elements as a function of their charge and spin. The kind of correlation found would require more extensive experimental data to establish it on a firm basis, and it must therefore be regarded as tentative.

The nuclei, whose spins and magnetic moments are listed in various tables (Breigleb, 1940; Mattauch, 1942; Kellogg and Millman, 1946), can be divided into four types dependent on whether the number of protons (p) and neutrons (n) they contain is even or odd. It is then found, as is well known, that the spins and magnetic moments are different for the different types as shown in table 1.

Table 1

Nuclear type	No. of protons and neutrons	Spin	Value of moment (nuclear magnetons)	Dependence on spin
A	$2p$ $2n$	0*	0*	—
B	$2p$ $2n+1$	half integral	-1 to $+1$ approx.	no effect
C	$2p+1$ $2n$	half integral	$-\frac{1}{2}$ to $+6$ approx.	increases with spin
D†	$2p+1$ $2n+1$	integral	$+0{\cdot}4$ to $+0{\cdot}9$ approx. for light nuclei	—

§ 2. EMPIRICAL RELATIONS

Plots of the nuclear magnetic moment against spin for nuclei of types B and C have previously been made, as for example, by Way (1932), Wigner and Feenberg (1941) and Margenau and Wigner (1940), the results being as shown in figures 1 and 2. It will be seen that for each type the points lie near to two parallel curves, though not exactly on them. The curves of course have no existence between the integral or half integral values of the spin coordinate but are useful in displaying the variation with spin. Nuclei with an odd neutron (B) give curves

---

* Either zero or presumed zero because of absence of hyperfine structure.

† Only four light stable nuclei of this type are known, and two heavier nuclei $^{40}_{19}$K and $^{176}_{71}$Cp.

staying more or less horizontal, while for nuclei with an odd proton (C) the curves rise with increasing spin. The points for very light nuclei of $Z \leqslant 9$ do not lie on the curves and have not been plotted, but all these omitted points are given as tables on the two figures. The deviations from the curves for the heavier nuclei are in many cases well outside experimental error but rarely large enough to be appreciable compared with the separation of the curves.

If now the nuclei are displayed on a diagram whose coordinates are spin and nuclear charge Z (or number of protons) it is seen (figure 3) that there are regions within which all the points belong either to the higher or to the lower curve. The boundaries are not well defined due to the small number of points available, but the general effect is clear.

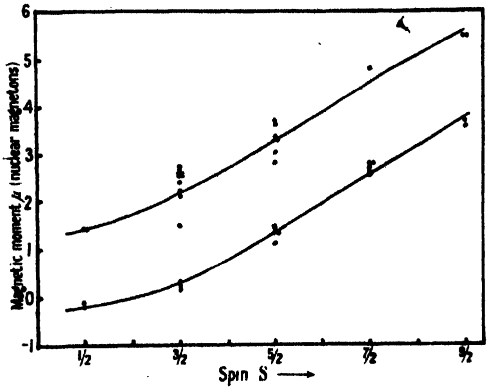


Figure 1. Nuclei of type C $(2p+1, 2n)$.
Points omitted.

	Spin	μ
$^{1}_{1}$H	1/2	2·789
$^{7}_{3}$L	3/2	3·253
$^{11}_{5}$B	3/2	2·686
$^{15}_{7}$N	(?)1/2	−0·280
$^{19}_{9}$F	1/2	2·625

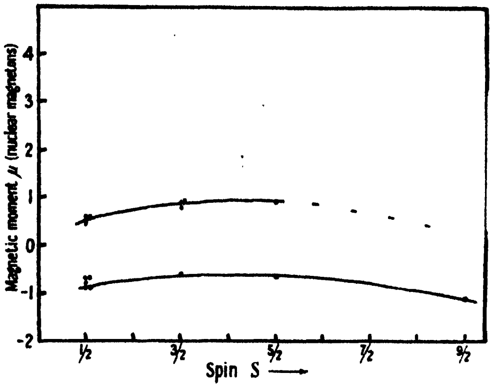


Figure 2. Nuclei of type B $(2p, 2n+1)$.
Points omitted.

	Spin	μ
$^{1}_{0}$N	1/2	−1·935
$^{9}_{4}$Be	(?)3/2	−1·176
$^{3}_{6}$C	1/2	+0·701

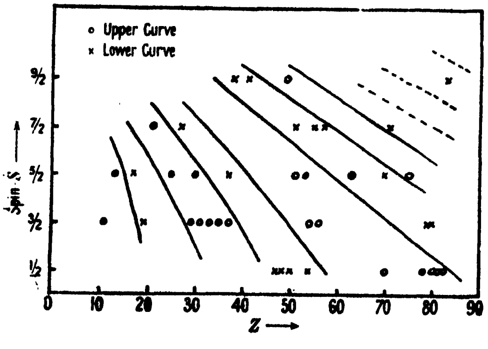


Figure 3.

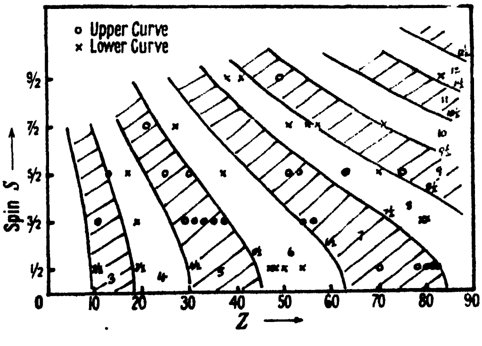


Figure 4.

In trying to express the limits of the regions mathematically some choice must be made, and the following method was adopted. Multiplication of Z by a factor varying with S allows the regions to be made into vertical bands. The appropriate factors are found to be proportional to $(1+\alpha S^2)$ where α is 0·071 as shown in table 2. The limits then occur at values of the quantity $Z(1+\alpha S^2)$ given in table 3, where thay are compared with $1{\cdot}50(k+\frac{1}{2})^2$, k being an integer varying from 3 to 9.

The two factors together would give lines of demarcation dependent on $\frac{1{\cdot}50\,(k+\frac{1}{2})^2}{1+0{\cdot}071\,S^2}$, and these are superimposed on the experimental data in figure 4. It must of course be pointed out that neither the form of the variation with S nor that with k is unique, for example, a dependence on $Z(1-0{\cdot}122\,S)$ would also allow for the effect of spin. The point at $Z=63$ is due to $^{151}_{63}$Eu and $^{153}_{63}$Eu, which have a spin of 5/2 and lie on the upper and lower curves respectively.

Table 2

Spin	$(1+0{\cdot}071S^2)$
0	1·00
1/2	1·02
3/2	1·16
5/2	1·44
7/2	1·87
9/2	2·44

Table 3

$Z(1+\alpha S^2)$	k	$1{\cdot}50(k+\frac{1}{2})^2$
$\simeq$ 19	3	18
$\simeq$ 29	4	30
47	5	45
$\simeq$ 60	6	63
$\simeq$ 85	7	84
105	8	108
$\simeq$ 125	9	135

These results suggest that the nuclear magnetic moment for zero spin lies on the upper or lower curve if $c\sqrt{Z}$ (c being a constant) is nearest to an odd or even integer respectively, the odd number giving the higher moment. On the earlier theories of nuclear structure (Bethe and Bacher, 1936), some such dependence on a quantum number is not unlikely. It is also possible that if the spin is partly due to the rotation of the nucleus it may modify the number of particles in a given level.

Figure 5 shows the neutron excess (A-$2Z$) plotted against Z for naturally occurring nuclei of type A, where the spin is probably zero, and below on the same diagram the values of Z for $S=0$, for which the magnetic moment is changing over. The correlation suggests that a lower neutron excess occurs when the change-over is taking place.

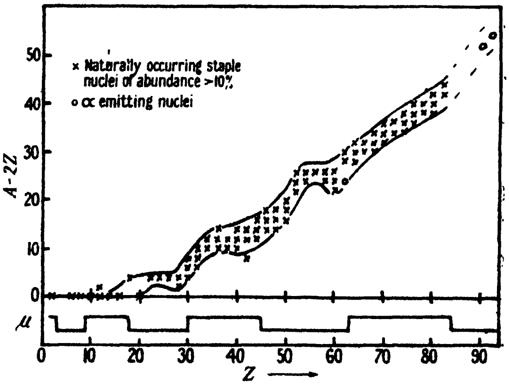


Figure 5.

It would be unwise at present to base much on the degree of agreement shown in figure 4, but if the division into regions could be well established and their boundaries defined more exactly, a useful indication might be obtained as to the kind of structure which exists in the heavier nuclei.

REFERENCES

Bethe and Bacher, 1936. *Rev. Mod. Phys.*, **8**, 171.
Breigleb, G., 1940. *Atome und Ionen* (Leipzig).
Kellogg, G. and Millman, 1946. *Rev. Mod. Phys.*, **18**, 348.
Margenau and Wigner, 1940. *Phys. Rev.*, **58**, 103.
Mattauch, J., 1942. *Kernphysicalische Tabellen.*
Way, K., 1939. *Phys. Rev.*, **55**, 963.
Wigner and Feenberg, 1941. *Reports on Progress in Physics* (*Phys. Soc.*), **8**, 294.

THE MATCHING OF HIGH-FREQUENCY TRANSMISSION LINES USING A FREQUENCY-VARIATION METHOD

By A. S. EDMONDSON,
The Inter-Services Research Bureau;
now at the Sir John Cass Technical Institute, London

MS. received 12 *March* 1947

ABSTRACT. A frequency-variation method which shows when an impedance is correctly matched to a transmission line is described. The variation with frequency in the sending-end voltage of a transmission line many wavelengths long is reproduced on the screen of a cathode-ray oscilloscope and from the curve it is possible to see whether the standing waves along the line are large or small in amplitude, and how they vary with frequency. A description is given of an apparatus applying the new method to ultra-high frequencies and used in connection with the matching of aerials and filters to a coaxial transmission line. The method has also been used for demonstrating several important transmission line properties and in this connection is useful for educational purposes.

§ 1. THE FREQUENCY-VARIATION METHOD

STANDING waves along a transmission line of finite length occur when the line is terminated in an impedance other than its characteristic impedance; they are caused by interference between the incident wave and the wave reflected from the terminating impedance, and can be detected by measuring the variation in the sending-end voltage V_s, as the electrical length βl of the line is altered. Neglecting attenuation, typical curves showing the relation between V_s and βl are shown in figure 1. (l is the physical length of the line, $\beta = 2\pi/\lambda = 2\pi f/v$, where f = the frequency, v = the velocity of propagation along the line and λ = the wavelength measured along the line.)

Figure 1 (*a*) is the general case of a loss-free line terminated by an impedance Z_T having both resistive and reactive components. The standing-wave ratio ($V_{\max}/V_{\min}$) depends on the ratio Z_T/Z_0, where Z_0 is the characteristic impedance of the line. Figure 1 (*b*) is the case of the line open- or short-circuited, or terminated by a reactance only; the shapes of the curves are the same for each, but the distances of the maxima and minima from the end of the line are different. The standing-wave ratio is theoretically infinite. Figure 1 (*c*) shows the relation when the line is terminated by its characteristic impedance; there are no standing waves and the voltage is independent of the electrical length of the line. In all practical cases, attenuation causes the incident and reflected waves to decrease in amplitude exponentially with distance along the line, resulting in a diminution of the standing-wave ratio as the distance from the termination increases.

The electrical length of the line cannot be altered satisfactorily by decreasing l owing to the practical difficulty of increasing it again; it is therefore changed by varying the frequency. It is shown later that the line must be long in order that the curve showing the variation of voltage at the sending end with frequency, the (V_s, f) curve, should have several maxima and minima when standing waves are present. The general type of curve obtained when the line is terminated by an impedance Z_T is shown in figure 2(*a*). The exact positions of the maxima, as well as their number, depend on the length of the line; the broken line shows the envelope of the series of curves that would be obtained if l were increased or decreased gradually by an amount $\lambda/2$. When Z_T is constant, the general shape of the curve is the same over the frequency range, but when Z_T varies, the relative amplitudes of the voltage maxima alter accordingly. If the range of frequencies includes a region in which the value of Z_T approaches the characteristic impedance of the line, the envelope of the (V_s, f) curve shows a minimum as in figure 2(*b*), since then the standing-wave ratio is least. If Z_T becomes equal to Z_0, the range of frequencies for which this condition is fulfilled gives no variation of V_s and the envelope of the curve crosses over where matching occurs, as in figure 2(*c*).

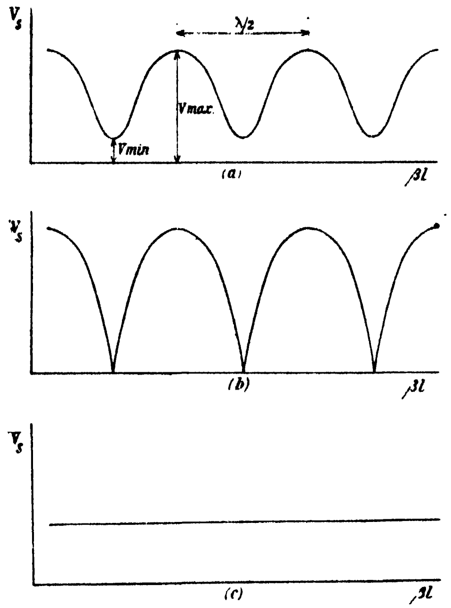


Figure 1. The variation of the sending - end voltage V_s of a transmission line with the electrical length βl;

(*a*) when the line is terminated in an impedance $Z_T = R_T + jX_T \neq Z_0$;

(*b*) when the line is terminated in an open- or short-circuit or a pure reactance;

(*c*) when the line is terminated in its characteristic impedance Z_0.

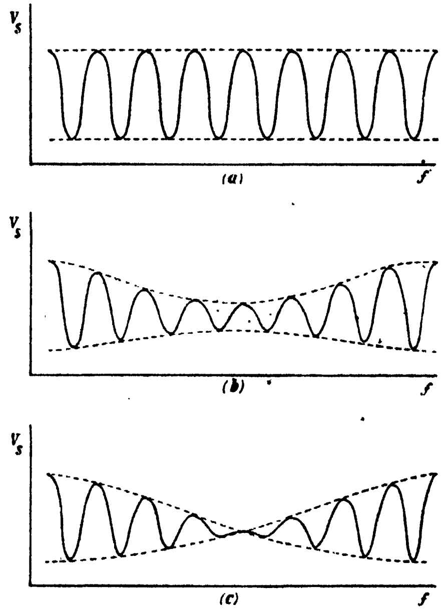


Figure 2. The variation of the sending-end voltage V_s of a transmission line, many wavelengths long, with frequency f;

(*a*) when $Z_T \neq Z_0$ and does not vary appreciably with frequency;

(*b*) when Z_T approaches Z_0 at a frequency within the range of observation;

(*c*) when $Z_T = Z_0$ at a frequency within the range of observation.

So that changes in the standing waves along the line due to a small change of Z_T may be readily detected on the (V_s, f) curves, the maxima must be closely spaced, i.e. $\Delta N/\Delta f$ should be as large as possible, ΔN being the number of waves appearing in the (V_s, f) curve in the frequency range Δf. Since the maxima of the standing waves along the line are spaced at intervals of $\lambda/2$, the number of half-wavelengths, N, in the line measures the number of standing waves at a frequency f, i.e.

$$N=\frac{l}{\lambda/2}=\frac{2lf}{v}.$$

The change in frequency Δf producing a change in N of ΔN is given by the equation

$$\Delta N=\frac{2l}{v}\Delta f.$$

But since ΔN is also the number of waves in the frequency range Δf on the (V_s, f) curve,

$$\frac{\Delta N}{\Delta f}=\frac{2l}{v}.$$

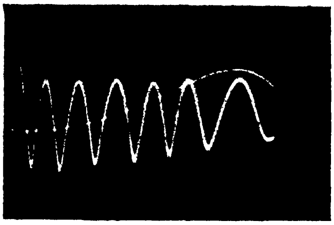
For this to be large the line must be long, but l must not be so large that the standing wave ratio and the variation of voltage at the sending end of the line are diminished too greatly on account of attenuation.

§2. DESCRIPTION OF THE APPARATUS

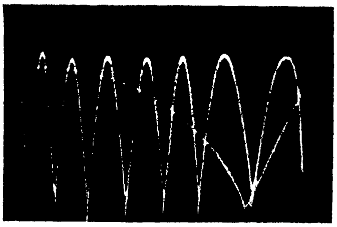
The instrument incorporating the new method was developed to show directly whether or not a piece of equipment was correctly matched to a coaxial cable, and, in particular, to enable the results of adjustments to be seen immediately without any lengthy calculations or the need for knowing the impedance of the line or the termination. Elimination of standing waves was important, at the time, not so much from the point of view of reducing losses, but to prevent changes in the sending-end impedance for different lengths of cable. Provided that the terminating impedance does not vary too rapidly with frequency, the present method is particularly suitable for this purpose.

A block diagram of the apparatus is shown in figure 3. The oscillator was fitted with a rotary condenser driven by a synchronous motor to vary the frequency. The output of the oscillator was taken to the long transmission line, connected at its far end to the terminating impedance, while the high-frequency voltage variations at the sending end of the line were rectified by a detector unit. The low-frequency voltage variations from the detector were amplified and applied to the Y-plates of the cathode-ray oscilloscope. The linear timebase of the oscilloscope was synchronized to the rotary condenser through a phase and amplitude control unit.

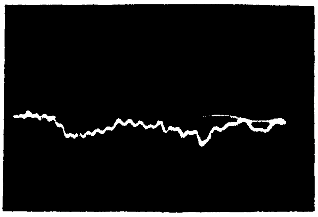
The oscillator was driven by two triode valves working in push-pull, and was tuned by a $\lambda/4$ resonant line short-circuited by a sliding bar which altered the mean frequency from about 280 to 400 megacycles per second. For a fixed setting of the shorting bar the oscillator was tuned over a range of about 15 megacycles per second by the rotary condenser, which was of the series-gap type and altered the capacity at the open end of the resonant line. Variable resistors in the anode and grid circuits enabled the oscillator to be operated under suitable conditions. The frequency of the ocillator was determined mainly by the position of the shorting bar and the setting of the variable condenser, but was affected slightly by the loading. The exact frequency for a given set of conditions was measured with an absorption wavemeter employing a tuned coaxial line resonator and a silicon-tungsten crystal, with a sensitive meter to indicate resonance.



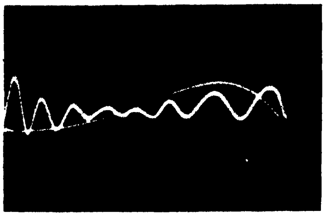
No. 1.



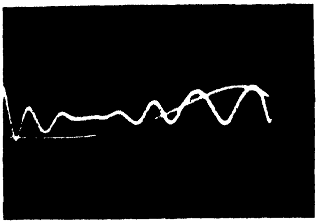
No. 2.



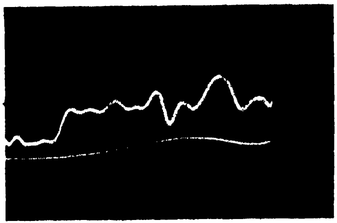
No. 3.



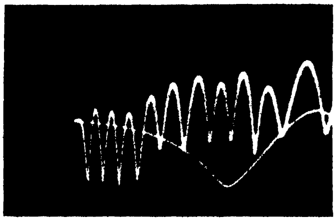
No. 4.



No. 5.



No. 6.



No. 7.

The transmission line was a coaxial cable (Telcon PT29M), consisting of a single inner conductor of copper insulated by polythene from the outer conductor of copper braiding. The characteristic impedance was 72 ohms and the attenuation, at 300 megacycles per second, was about 0·13 decibels per metre. The ratio (wavelength along the cable)/(wavelength in air) had a value approximately equal to 2/3. The most suitable length of cable depended on the type of experiment being performed, but about 50 metres, for which $\Delta N/\Delta f = 0{\cdot}5$ cycles per Mc./sec., was satisfactory for many applications in the present frequency range.

When a constant impedance load equal to the characteristic impedance of the cable was required a very great length was used, since, owing to the attenuation of the reflected wave, the standing-wave ratio at the sending end was small even when the line was not correctly terminated. Z_s therefore remained substantially constant and equal to Z_0 over the whole frequency range. In practice, using a cable 185 metres long terminated in a resistance approximately equal to Z_0, the variation in voltage observed at the sending end as the frequency was altered was very small.

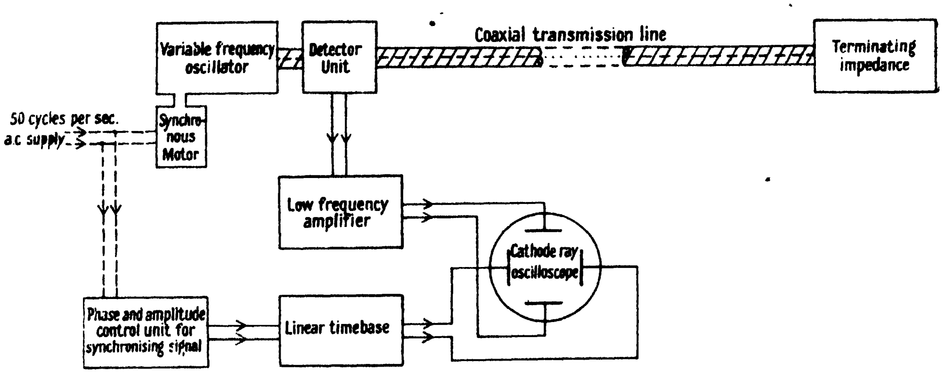

Variable frequency oscillator
Detector Unit
Coaxial transmission line
Terminating impedance
50 cycles per sec. a.c supply
Synchronous Motor
Low frequency amplifier
Cathode ray oscilloscope
Phase and amplitude control unit for synchronising signal
Linear timebase

Figure 3. Block diagram of the apparatus.

In using the apparatus qualitatively, i.e. simply to detect the presence of standing waves or to compare their relative magnitudes under different conditions, variations only of the high frequency voltage needed to be recorded; precautions necessary to measure the absolute voltages at these frequencies were therefore not required. The rectifier, using either a crystal or a diode valve, was built into a detector unit, which, to measure the voltage variations, was connected across the cable at the sending end. The output was applied to the Y-plates of the oscilloscope through the low-frequency amplifier. The curve obtained on the screen showed the variation of the sending-end voltage of the line with frequency.

§ 3. EXPERIMENTAL RESULTS

Photographs of the oscilloscope traces obtained in different cases are shown in nos. 1–7 of the plate.

No. 1 shows the general case of a cable, 43 metres long, terminated in an impedance Z_T. The voltage variation is similar to that shown in figure 2 (*a*).

No. 2 is the oscillogram for the same length of cable short-circuited at its termination. It has the same shape as the curve in figure 1 (*b*) and shows that the sending-end voltage varies over comparatively wide limits.

No. 3 was obtained using a cable 185 metres long terminated in a resistance approximately equal to Z_0. The small variations in voltage due to the very great length are seen, showing that the sending-end voltage of the line remained fairly constant. The larger variations were due to irregularities in the cable which are explained later; because of these irregularities, this sample of cable could not be used as a constant load for experimental purposes, but, at the time the photographs were taken, it was the only really long piece available.

Nos. 4 to 7 are explained in the following applications which demonstrate the usefulness of the new method.

Matching a dipole aerial to a cable

A dipole aerial, specially made for demonstration purposes, was constructed so that both the resistive and reactive components of its impedance could be adjusted over wide ranges, thus enabling it to be matched to the cable at any frequency within the range of the oscillator. The dipole arms were fixed at one end of a pair of parallel lines acting as a matching transformer, while at the other end was a shorting bar, adjustable in position. The cable was connected across the matching transformer between the dipole arms and the shorting bar, and, by adjusting the position of the connecting point, both the resistance and the reactance at the end of the cable could be altered, while moving the shorting bar varied the reactance only. (A balancing transformer consisting of a coaxial sleeve approximately $\lambda/4$ in length coupled the unbalanced cable to the balanced matching transformer.)

When the aerial was connected to the apparatus through a cable 43 metres long, the curve obtained on the screen of the oscilloscope was as shown in no. 1. The frequency range observed at any instant was about 15 Mc./sec., but, by altering the mean frequency of the oscillator, the variation of V_s was followed over the whole range from 280 to 400 Mc./sec.

The aerial was matched to the cable at a particular frequency in the following way. The required frequency was brought to a known point on the screen by adjusting the mean frequency of the oscillator, and the positions of the cable connecting point and the shorting bar on the matching transformer were then adjusted until a minimum in the voltage variation was observed as in no. 4. It was found that this could soon be done, and no difficulty was experienced in obtaining the necessary conditions. By means of further small adjustments a curve was obtained having a flat portion, indicating the disappearance of standing waves along the cable and showing that matching occurred at that point. The point at which matching occurred was then brought to the required position on the frequency axis by making suitable adjustments to the aerial. No. 5 shows the matching curve finally obtained.

The effect of reflection of the radiation from the aerial by a neighbouring object was observed by the change in the curve on the oscilloscope screen. A reflector was brought slowly up to the arms of the aerial originally matched to the cable feeding it. The changes in impedance so caused had an effect on the standing waves along the cable, and periodic variations in the (V_s, f) curve were observed, which increased greatly in magnitude as the reflector approached the aerial. For a given position of the reflector, in order to match the aerial to the cable again, further adjustments of the cable connecting point and the matching transformer shorting bar had to be made.

Adjustment of ultra-high frequency filters

In a particular application it was required to short-circuit one frequency f_1 without causing any mismatch to a second frequency f_2. If λ_1 and λ_2 are the wavelengths corresponding to the frequencies f_1 and f_2 respectively, an open-circuited stub of electrical length $\lambda_1/4$ will short-circuit f_1 when placed across a line. The reactance of this stub at the frequency f_2 can be tuned out with a second open-circuited stub whose electrical length is $(\lambda_2/2-\lambda_1/4)$ without affecting its short-circuiting properties at the first frequency. Figure 4 (*a*) shows the arrangement of the stubs across the line.

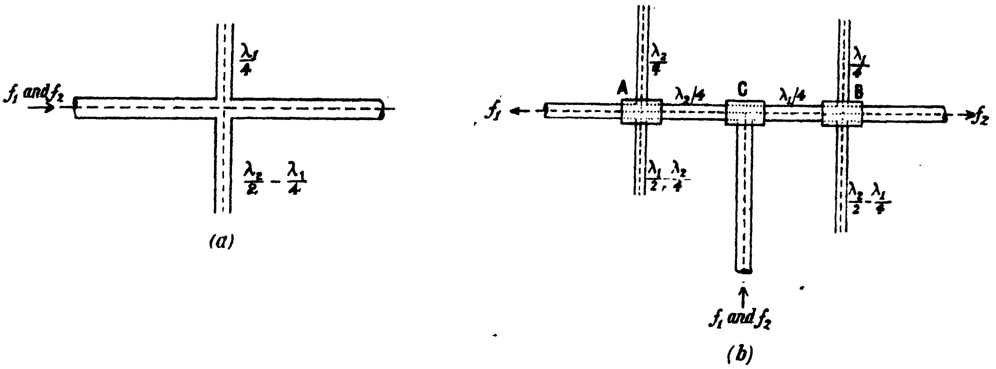


Figure 4. High-frequency filters.
(*a*) Filter short-circuiting f_1 and matched to f_2.
(*b*) Filter separating two frequencies f_1 and f_2.

A junction box was constructed, having two coaxial side-plugs for connecting the stubs across a coaxial cable. The stubs themselves consisted of short lengths of coaxial cable attached to sockets which fitted on to the side-plugs of the junction box. The electrical length of a stub could be altered by shortening, but for experimental purposes it was more convenient to connect a small vaaiable condenser to the open end to give a continuous adjustment over a small range.

To one end of the junction box was connected the usual long piece of cable leading to the oscillator and to the other end a very long piece of the same type of cable was connected to provide a constant impedance load. To short-circuit the first frequency, a stub was fitted to one side-plug, and a detector unit, whose amplified output was applied to the Y-plates of the oscilloscope, was fitted to the other. The curve obtained on the screen showed the variation with frequency of the voltage across the line at the junction; the frequency at which a short-circuit was produced was shown by a pronounced minimum on the curve. The electrical length of the stub was adjusted until the minimum occurred at the required frequency f_1. To tune out the reactance of the first stub at the second frequency, the detector unit was placed at the oscillator end of the cable, to measure V_s, and a second stub was fitted to the second side-plug. The output of the detector unit was applied to the oscilloscope and the usual matching curve was obtained. By adjusting the electrical length of the second stub, matching was achieved at the frequency f_2. In contrast to the relatively wide flat portion on the matching curve in the case of the dipole aerial, matching occurred at a single well-defined frequency in the case of a filter, and the stub lengths had to be correct to within close limits.

A further problem in connection with filters, which was solved with the aid of the new method, was the separation of two frequencies f_1 and f_2 without causing any mismatch to either. One filter, A, similar to that described above, was adjusted so that it was matched to f_1 while short-circuiting f_2, and another filter, B, was matched to f_2 while acting as a short-circuit to f_1. A was connected to a junction box C, see figure 4 (*b*), through a cable of electrical length $\lambda_2/4$, and B was connected to C through a $\lambda_1/4$ length of cable. The short-circuit at A appeared at C as an open-circuit, thus causing no mismatch to f_2 which passed through B unaffected. Similarly the short-circuit at B for f_1 caused no mismatch at C for that frequency which therefore passed through A. The $\lambda/4$ lengths of cable between the filters and the junction box were calculated as accurately as possible before assembly, but their final values were those which gave the best results for matching. When the lengths were correct and C was connected to the oscillator through a long cable, the (V_s, f) curve showed two sharp minima corresponding to matching at f_1 and f_2, although they did not both appear on the screen simultaneously owing to the difference in the two frequencies. Any deviation from the $\lambda/4$ lengths for the connecting cables was detected by imperfection in the matching curve.

Results obtained with faulty samples of cable

With certain samples of cable the usual flat portion of the matching curve was not obtained, even though it was known that the line was correctly terminated. There was a general flattening of the curve but it was distorted as though the usual curve were superimposed upon another. An example of this is shown in no. 6, which was the closest approach to the usual matching curve that could be obtained in a particular case. No. 7 shows the same cable, 76 metres long, short-circuited at its end. These photographs should be compared with nos. 5 and 2 respectively, which show the corresponding normal curves for a shorter cable. The explanation appears to be that somewhere along the cable the electrical properties changed, causing a region of mismatch; reflection taking place from this region would alter the form of the standing waves at the sending end of the cable and even when the cable was matched at its far end the standing waves would not disappear. The larger variations appearing in no. 3 for a very long cable, and mentioned earlier, were of the same nature and must have been due to faults in the cable.

It is interesting to note here that an alternating-current method of fault location in telephone cables has been described by Palmer and Tufrail (1930) in which the sending-end impedance of a cable is measured at audio frequencies using a bridge circuit. When a fault exists at a single point along the cable, the sending-end impedance shows the same type of variation as in figure 2 (*a*) and the distance of the fault is calculated from the spacing of the maxima. Complications arise when more than one fault is present, as the several reflected waves give rise to such an intricate pattern that the simple calculation can no longer be made. It is, of course, necessary to use audio frequencies for cables of this nature, and on this account the distances involved are greater than at higher frequencies. Roberts (1946) has discussed theoretically both pulse and frequency modulation methods for locating cable faults, with particular reference to high frequencies. He notes that considerable errors may arise using the latter method when more than one fault is present, and concludes that the pulse method can be applied more readily to cable-fault location than a frequency-modulation system giving equal information about the amplitudes and positions of the faults. The paper, which is followed by a comprehensive discussion, may be consulted for a mathematical treatment.

§4. SUMMARY

The frequency-variation method of matching described in this paper has, so far, only been used qualitatively, but it has proved extremely helpful in the adjustment and testing of various types of high-frequency apparatus for which matching is required. The new method is useful in that it will show when and at what frequency and over what range the line is correctly terminated, in addition to which it indicates directly on the screen of an oscilloscope, the relative magnitudes of th standing waves at the sending end of the line over a wide frequency range, and shows visually the effect of adjustments to the impedance terminating the line. Its use in the demonstration of transmission-line properties should be noted, the great advantage here being that, as various adjustments are made to the line and its termination, the instrument gives a continuous visual record on the screen of the oscilloscope. An added advantage is that, in most cases, the results are easily interpreted. The method can be further developed by calibration to enable quantitative results to be obtained; in conjunction with the usual impedance measuring instruments, this would give a fresh approach to certain transmission-line problems.

The application of a similar method to audio frequencies, in connection with fault location in telephone lines, has already been mentioned; at higher frequencies than those employed in the present investigation, where the physical length of line needed would be smaller, it should prove to be as useful as in the applications described in this paper, particularly in the matching of waveguides. The same method can be extended to waves other than electromagnetic waves, e.g. longitudinal sound waves, the requirements being simply a variable frequency source, a "transmission line" and a detector.

ACKNOWLEDGMENTS

I wish to express my gratitude to Dr. N. L. Yates-Fish for his help and advice while the experiments were being performed and for originally suggesting the problem to me.

The work was carried out for the Inter-Services Research Bureau, to whom I am indebted for permission to publish this paper.

REFERENCES

Palmer, W. T., and Tufrail, M. E., 1930. *J. Post Office Elect. Engrs.*, **23**, 42.
Roberts, F. F. 1946. *J. Instn. Elect. Engrs.*, **93**, III, 385.

THERMOELECTRIC POWER OF CADMIUM OXIDE

By J. P. ANDREWS,
Queen Mary College, London

MS. received 28 *January* 1947

ABSTRACT. The thermoelectric e.m.f. (E) of cadmium oxide CdO against platinum is measured over a temperature range of −110° to 800° c. The results above 100°c. are well represented by the formula $E=-aT+b\log_e T-C$, in which the values of the constants a, b, and c vary somewhat according to the previous treatment of the specimen ; their values in a number of different experiments are given. When E is in millivolts, their mean values are $a=0{\cdot}154$, $b=34{\cdot}3$, and $c=225$. The results are discussed, and if it is assumed that the number of conduction electrons is given by $n=Be^{-\beta/kT}$, β is found to be 0·034 electron volts. The sign of the thermoelectric power indicates that CdO is an excess conductor.

§ 1. INTRODUCTION

SEMI-CONDUCTORS, of which CdO is an example, commonly exhibit large thermoelectric (e.m.fs.) against a pure metal; and if this property is associated with low resistance, as it is in CdO, measurements of thermoelectric power can be made by the ordinary methods applicable to metals. These measurements yield information about the nature of the semi-conductor, and of the current carriers to which the electrical conductivity and thermoelectric properties are due. In terms of the accepted terminology of the subject, this information obtainable may be summarized thus:—

(i) The algebraic sign of the thermoelectric power indicates whether the semi-conductor is to be regarded as of "defect" or "excess" (or alternatively "abnormal" or "normal") type. The conductor will be of normal or excess type if the current is carried mainly by electrons; and will be of abnormal or defect type if conduction is mainly due to "holes" which behave like positively charged particles. It is worth noting at this point that in a semi-conductor whose current is carried by both electrons and holes, another distinction arises. According to Fowler (1933) a semi-conductor of this kind which is normal at low temperatures retains a negative thermoelectric power at high temperatures; whereas the thermoelectric power of a semi-conductor which is abnormal at low temperatures will change from a positive value at low temperatures through zero to a negative value at high temperatures. Above a certain *neutral temperature*, then, the sign of the thermoelectric power would indicate an excess or normal conductor.

(ii) In simple instances, where only one type of carrier need be considered, the thermoelectric power of a semi-conductor against a metal is given (Wilson, 1939) by

$$\frac{dE}{dT} = \mp \frac{k}{e} \log \frac{n_2}{n_1} \qquad \ldots\ldots(1)$$

and it is often permissible to ignore the specific contribution of the pure metal In this formula n_2, n_1 are the numbers of electrons or "holes" per unit volume taking part in the process in the semi-conductor and metal respectively, a negative sign being associated with normal or excess conductors.

The equation is of the type derivable from classical theory and is an example of the sort of simplification occurring in semi-conductors, where the number of carriers is so small that the electron gas responsible for the conduction may be considered non-degenerate and having a Maxwell distribution of velocities. It has to be applied to actual semi-conductors with some reserve, since in the first place it implies only one kind of carrier, and in the second place is strictly applicable to metals in which n and n_2 do not vary appreciably with temperature, whereas a fundamental feature of semi-conduction is the increase of n_2 with increase of temperature. However, if the law of intermediate temperatures applies to a thermoelectric circuit containing a semi-conductor as one material, then the change of e.m.f. dE round the circuit when the temperature of the hot junction is raised from T to $T+dT$ (the cold junction being kept at constant temperature) is equal to the e.m.f. in the circuit when the hot junction is at $T+dT$ and the other at T, the mean temperature in this case being very close to T throughout. According to this argument, therefore, the gradient of the e.m.f./temperature curve gives the thermoelectric power which would be obtained from a thermocouple substantially at a temperature T, and the value of n_2 resulting from equation (1) is the electron density in the semi-conductor at a temperature T.

Some of this information is obtainable from other measurements (e.g. of conductivity and Hall effect), which are often preferred to measurements of thermoelectric power because the interpretation offered by current theory appears to be easier, particularly simple relations connecting them with carrier density and charge. It may be, however, that this impression is rather illusory, at any rate where samples in the form of powders are used; for then the shape of the individual grains enters the calculation of both conductivity and Hall effect. This fundamental difficulty is generally ignored, perforce; nevertheless, even in compressed samples, a large uncertainty must remain. Thermoelectric measurements on the other hand are likely to be free from this trouble; and for materials like CdO, the measurements certainly are easier and more accurate than measurements of the Hall effect.

The principal work already carried out on the thermoelectric properties of CdO includes the experiments of Bädeker (1907) on very thin films and over a limited temperature range, and that of Fischer, Dehn and Sustman (1932) for a single temperature difference between hot and cold junctions. Bädeker's experiments yielded a value of 30 to 35 microvolts per degree for the thermoelectric power, while Fischer, Dehn and Sustman (1932) gave 84 microvolts per degree for junction temperatures of 9° and 595° c. Experiments on the allied topic of electrical conduction were carried out by Baumbach and Wagner on bulk material in 1933.

§ 2. MATERIALS

Two varieties of material were used, after the manner of Baumbach and Wagner. These were:—

(*a*) CdO powder, the purest obtainable. A chemical analysis carried out on this powder revealed the following quantities of small impurities:

Lead	0·022 %
Iron	0·009 %
Zinc	0·016 %
Copper	Trace

This material is of a light chocolate brown colour, which darkens temporarily at high temperatures.

(*b*) $CdCO_3$ powder, which, when made into a paste with water, can be moulded into appropriate shapes, dried and calcined, when the brown oxide CdO is obtained. Considerable reduction of volume occurs on heating the white carbonate specimen, and it was in practice impossible to prevent distortion and fissures. The resulting sample of CdO is quite hard and coherent and has a density of about 4·9 gm./c.c., the density of solid CdO being given as 8·1 in the International Critical Tables.

§ 3. METHOD OF MEASUREMENT

Of the methods used, the most fruitful was the following (see figure 1). The specimen in the form of a rod or block, or else of a more or less loose powder contained within a porcelain collar, was placed upright on a thin platinum disc, which in turn was silver-soldered to a brass block on which an electric furnace was wound. The brass block was bored to take the leads and protecting sheath of a chromel-alumel junction which was silver-soldered on to the lower surface of the platinum disc, after passing through a small aperture in the top of the brass block. On top of the specimen, an exactly similar platinum disc was pressed, the disc forming the bottom of a vessel in which cold water circulated. Platinum wires from each disc lead to a potentiometer.

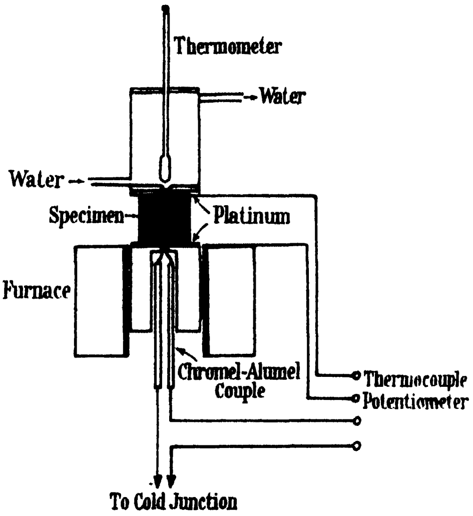


Figure 1.

The chromel-alumel thermo-junction, by which the temperature of the hot end of the specimen was to be measured, was carefully calibrated at a number of fixed points between 0° c. and the melting point of NaCl.

§ 4. RESULTS

In all experiments above room temperature the results resembled those illustrated in the curves of figure 2, but considerable variation occurred between different specimens and for different heat treatments.

In order to summarize the results the curves were fitted to an empirical equation $E = -aT + b.\log_e T - c$, from which

$$\frac{dE}{dT} = -a + \frac{b}{T}.$$

Values of the constants a, b, and c in this equation are given in table 1 for various specimens subjected to different treatments. T is the absolute temperature of the hot end of the specimen.

It may be remarked that the equation

$$E = -aT + b.\log_e T - c$$

used for this summary is not the only type of equation to which the curves may be fitted over a considerable interval of temperature. It is nevertheless true that whether this formula is taken to represent the whole family of curves, or whether it

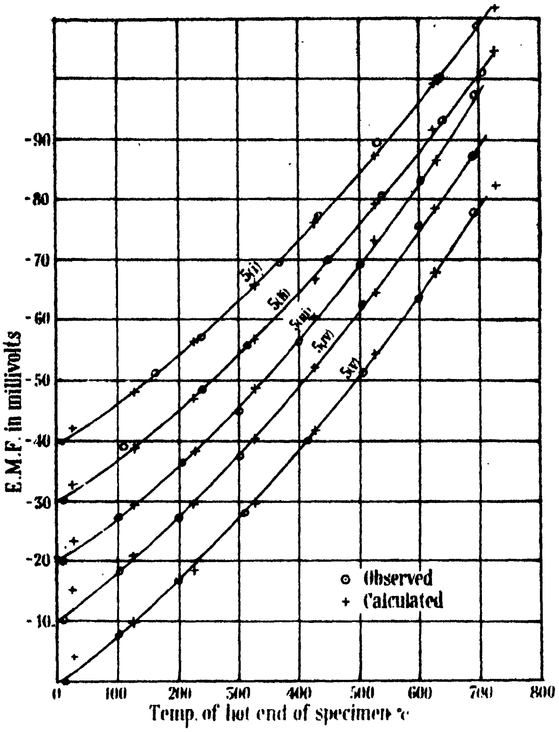


Figure 2. Results of experiment 5 of Table 1. The e.m.f. scale applies to the lowest curve, viz. 5 (v). For the other curves the origin should be raised successively in steps of 10 millivolts.

Table 1

Specimen	Temp. range (°c.)	a (volt/deg. $\times 10^3$)	b (volt $\times 10^3$)	c (volt $\times 10^3$)
1. Loose powder in glass tube	20–460	0·165	34·5	149
2. Stick, from $CdCO_3$, 3 cm. long, heated 4 hours at 800° c.	20–810	0·183	33·7	139
3. Flat pellet from $CdCO_3$	20–700	0·105	26·1	117
4. Same, repeated	20–700	0·114	28·5	128
5. Series of cylinders made from $CdCO_3$:				
(i) heated for 1 hour at 750° c.	20–750	0·155	32·7	140
(ii) ,, ,, 1·75 hours at 750° c.	,,	0·159	34·1	147
(iii) ,, ,, 2·5 ,, ,, ,,	,,	0·185	43·5	198
(iv) ,, ,, 3·25 ,, ,, ,,	,,	0·177	36·0	158
(v) ,, ,, 4·0 ,, ,, ,,	,,	0·170	35·4	150
6. CdO powder packed in a narrow silica tube	20–700	0·153	37·0	166
7. Loose CdO powder in porcelain collar	20–700	0·160	38·3	170
8. Same, repeated	20–700	0·146	31·6	137
9. Pellet, pressed with a hand-press from CdO, after grinding and heating the powder	20–735	0·132	28·9	126
10. Same, repeated	20–720	0·145	34·9	155
11. Machine pressed pellet, from CdO powder. Density 55% of solid density	20–750	0·160	39·7	176

is used for individual curves, it does appear to fit satisfactorily over a wider temperature range than several other 3-constant equations tried. It is chosen for that reason, and because there are some theoretical grounds for anticipating a relation of this character. The agreement between calculated and observed e.m.f. is sufficiently exemplified by the curves of figure 2, which are quite typical.

§ 5. DISCUSSION OF RESULTS

1. The empirical equation selected suggests that at a temperature $T_m = b/a$, the thermoelectric power becomes zero, and at lower temperatures positive. The interest in this extrapolation derives from the prediction of Fowler, mentioned above, that a semi-conductor whose thermoelectric power is positive at low temperatures, and in which carriers of both signs take part in the effect, ought to show just such a change from positive to negative thermoelectric power. From the values of a and b in table 1, the mean value of T_m should be 225° K., or −48° C.

Experiments were accordingly carried out to continue the curve below room temperature, by the following method (see figure 3). The specimen was firmly held between two platinum discs from which platinum wires led to the potentiometer. The upper disc was maintained at about room temperature by flowing water, while the lower disc was supported on a copper rod capped by a copper disc to bear against the platinum. A Dewar vessel containing liquid air was slowly raised until a chosen length of the rod was immersed. In this way the lower platinum disc was cooled by conduction to a temperature depending on the length of rod left exposed. The temperature was measured as before by a chromel-alumel junction soldered to the platinum disc. The calibration of the couple was checked by measuring the freezing point of absolute alcohol.

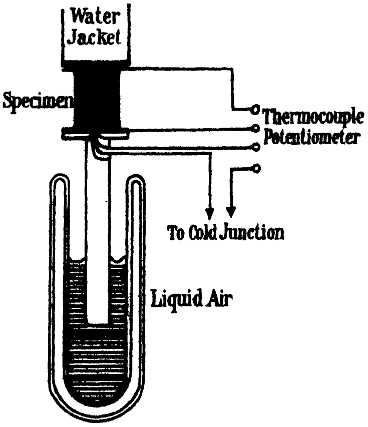


Figure 3.

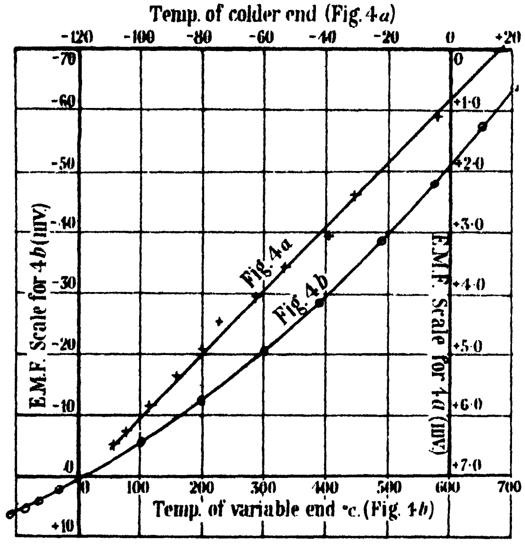


Figure 4. Continuation of the e.m.f. temperature curve to low temperatures.

The result obtained (see figure 4a) shows that over this restricted range of temperature, the e.m.f. is very nearly proportional to the difference of temperature between the hot and cold ends of the specimen. The complete curve for this specimen, from −110° to 700° c. is given in figure 4*b*.

Similar results were found for two other specimens.

2. It appears from the foregoing that the equation

$$\frac{dE}{dT} = -a + \frac{b}{T},$$

though applicable over a range of several hundred degrees, cannot be extrapolated to low temperatures. It may be, therefore, that the equation is to be regarded as purely empirical, without much physical significance. On the other hand, however, it is not uncommonly found that the electrons may be excited to the conduction band in a semi-conductor from two different energy levels. The number of conduction electrons may then be supposed given by an expression like

$$n = Ae^{-\alpha/kT} + Be^{-\beta/kT},$$

the activation energies α and β being different. In such a case the term with higher activation energy tends to take over at the higher temperatures, the other at the lower. It is possible that the higher term is dominant in the CdO results above 100° c., while the influence of the lower term begins to appear at lower temperatures (cf. Mott and Gurney, *Electronic Processes in Ionic Crystals*). In the experiments of Baumbach and Wagner (1933) there is indeed some indication of a change in the law of variation of conductivity with temperature, somewhere between 20° and 500° c., and the change is of the kind described. To settle the point for thermoelectric measurements requires experiments over a wider temperature range, but that there is evidence in the present results for a transition of the required type is shown in figure 5 where values of dE/dT are plotted against $1/T$ for three representative experiments. In practically all cases a fairly sharp change in the

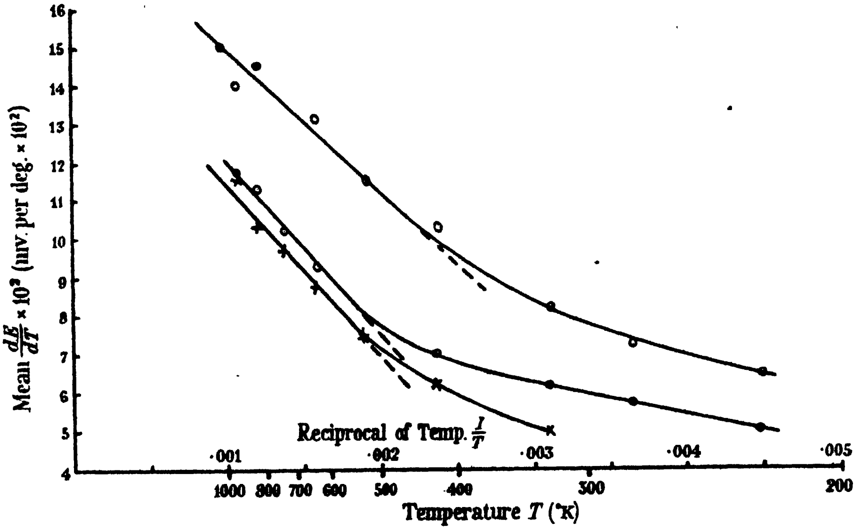


Figure 5. The marked points represent the measured mean values dE/dT over a range fo temperature, usually 50° either side of the temperature T, except at the lowest temperatures where half this range is taken.

slope of such curves appears between 400° and 500° K. In all cases the graphs are approximately rectilinear above that temperature. Assuming then a process like that described above, the value of n_2 in equation (1) may be put equal to $Be^{-\beta/kT}$. It then follows that $\beta = b \, . \, e$ ergs, where e is the charge on the electron, or alternatively, $\beta = b$ electron volts. From table 1, the mean value of b then gives the activation energy of β as 0·034 electron volts.

3. Over the temperature range $-120°$ to 800° C. the thermoelectric power is negative against platinum. CdO behaves therefore as an excess or normal semi-conductor.

4. While it is clear that the treatment of the specimen has a considerable effect on the values of the constants *a*, *b*, and *c*—an observation conforming to the usual and often-recorded experience in work on semi-conductors—it is not easy to lay down consistent rules. It is generally true that prolonged heating raises the values, though exceptions appear in table 1. In the series No. 5 of the table, it appears that the constants were increased when the specimen was heated up to a period of 2½ hours, but thereafter decreased. It is generally found that the higher the temperature of heating the more pronounced the effect, and as a rule all three constants increase and decrease together. After prolonged heating at a high temperature, the oxide tends to become darker, and although the darkening fades to some extent on cooling, there often remains a noticeable permanent effect. There is some evidence that a darker colour is associated with a greater stoichiometrical excess of cadmium (Cowley and Rees, 1946).

5. The state of compression or solidity of the specimen appears to have no consistent effect on the thermoelectric properties. This is to be expected, and constitutes the advantage of thermoelectric measurements over those which do depend on the state of aggregation of the specimen, as in experiments on conductivity.

ACKNOWLEDGMENT

For facilities for carrying out these experiments in his laboratory, and for helpful advice, my thanks are due to Prof. H. R. Robinson.

REFERENCES

BÄDEKER, K., 1907. *Ann. Phys., Lpz.*, **22**, 749.
VON BAUMBACH, H. H. and WAGNER, C., 1933. *Z. Phys. Chem.*, B, **22**, 199.
COWLEY, J. M. and REES, A. L. G., 1946. *Nature, Lond.*, **158**, 550.
FISCHER, F., DEHN, K. and SUSTMANN, H., 1932. *Ann. Phys., Lpz.*, **15**, 109.
FOWLER, R. H., 1933. *Proc. Roy. Soc.*, A, **140**, 505.
WILSON, A. H., 1939. *Metals and Semi-Conductors*, p. 56.

DISCUSSION

Prof. G. I. FINCH. I wish to draw attention to the possible effects of crystal structure and particularly crystal habit variations on the results. Cadmium oxide is of rock-salt structure and it is quite probable that compression, heating and so forth might well change the crystal habit in a micro-crystalline mass and hence affect the values of Dr. Andrews' constants.

Dr. H. H. HOPKINS. Perhaps we could be told whether Dr. Andrews is quite satisfied that the variations in the values of the constants of his equation do not derive from his method of "fitting" the equation to the experimental results, and that they do, in fact, indicate different physical conditions in the various samples under test?

Mr. J. H. Awbery. In fitting equations to his results, Dr. Andrews starts with the equation

$$\frac{dE}{dT}=a+\frac{b}{T}, \qquad \ldots\ldots \text{(i)}$$

of which the integrated form is

$$E=aT+b\,.\ln T+c, \qquad \ldots\ldots \text{(ii)}$$

where E is the thermal e.m.f. when the hot side of the specimen is at temperature T, the cold side remaining always at a temperature T_0.

In presenting the paper, he showed a slide of E versus T, with points calculated from the equation marked on it, and pointed out that in every case, the calculated point for the lowest temperature lay above the experimental curve. This is a serious matter, for it means that if the curve calculated from his equation were plotted, it would not give $E=0$ at $T=T_0$. This is clear, because according to (i) dE/dT is monotonic, and thus the curve representing (ii) cannot have a hump on it, which would correspond to a minimum of dE/dT. Without the hump, it cannot be correct at higher temperatures, high at lower ones, and correct at T_0. It follows that Dr. Andrews' calculated curve cannot be correct, and further, that the constants in the equation (ii) to this curve cannot be correct.

The error could arise if the thermometer with which he measured the temperature of the cold face did not agree with the thermocouple with which he measured that of the hot face, or if either did not in fact read the temperature of the face of the specimen, but a more likely explanation (especially in view of the author's skill as an experimenter) is that the calculated curve does not in fact properly represent the observations.

It was mentioned in the course of the discussion on the paper that the " calculated " curve was obtained by fitting an equation of the form (ii) by selecting three points (presumably on a curve drawn graphically). I would suggest that a better procedure, whenever there are conditions which a curve must satisfy, is to enforce compliance with these conditions first. Thus equation (ii) requires

$$0=aT_0+b\,.\ln T_0+c,$$

whence by subtraction

$$E=a(T-T_0)+b\,.\ln T/T_0 \qquad \ldots\ldots \text{(iii)}$$

or

$$\frac{E}{T-T_0}=a+\frac{b\,.\ln T/T_0}{T-T_0} \qquad \ldots\ldots \text{(iii } a\text{)}$$

In this form, the equation should give a linear plot of $E/(T-T_0)$ against $(\ln T/T_0)/(T-T_0)$, and the slope of the line should give b, its intercept giving a. c would then follow as $-(aT_0+b\,.\ln T_0)$. I suggest that by this procedure, all points could be given due weight, that the curve would necessarily comply with the condition at T_0, and that the values of a, b and c deduced would thus be more reliable.

Prof. R. W. Ditchburn. Do I correctly understand (a) that the curves were fitted by the method of least squares and (b) that the differences between samples are large compared with the difference between measurements on the same sample ?

Author's reply. I think it very likely that Prof. Finch is right, and that some part of the variation found ought to be ascribed to the processes he mentioned. The same is true of nearly all measurements on semi-conductors—and, I suppose, on solids in general.

I would like to refer Mr. Awbery to the paper itself, when a printed copy reaches him ; for I believe he will find his criticisms already dealt with there. In fact, the matter of the absence of a minimum in dE/dT, which he feels is so serious, is regarded in the same light in my paper, and was made a prominent part of my verbal account to the Society ; and I am grateful to him for adding such skilful re-emphasis. In the paper he will see that recognition of the failure of this test of the empirical equation led to further experiments and a suggested revision or restriction of the equation. Perhaps he would agree that under the circumstances, provided the equation is fitted as well as the curves show that it was, the particular method of arriving at the values of the constants is of secondary importance. That is not to say that his suggested method is ungratefully received, or will not be made good use of.

As Dr. Hopkins suggests, I daresay it is possible that with a different method of computation of the constants of the equation, there might be some slight difference in the variations I have given ; but the effects are large enough to leave no doubt of their existence.

I would not, however, be inclined to base any important conclusions on the numerical values given for these changes until more is understood about their origin, or about the right kind of experimental condition for their determination.

In reply to Prof. Ditchburn I may say that the method of least squares was not used. It is not easy to give a short answer to the second question, but I believe Prof. Ditchburn will be able to judge the relative differences on consulting the annotated table of results given in the paper.

THE CALCULATION OF POTENTIAL-ENERGY CURVES FROM BAND-SPECTROSCOPIC DATA

By A. L. G. REES,
Division of Industrial Chemistry, Council for Scientific and Industrial Research, Melbourne, Australia

MS. received 15 *January* 1947

ABSTRACT. Analytical expressions equivalent to the Rydberg-Klein graphical construction have been derived for the calculation of potential-energy curves of diatomic molecules from band-spectroscopic data. These expressions have been shown to lead to satisfactory values for D_e and r_e and, in contrast to the Rydberg method, enables the potential-energy (p.e.) curve to be evaluated with accuracy in the region of the minimum. The Morse function has been shown to satisfy the analytic expressions where a quadratic in $(v+\frac{1}{2})$ has been used to express the band data. A method for dealing with electronic states in which a discontinuity in the law of force occurs has been outlined. By using a series of quadratics to satisfy the band data, the calculation of p.e. curves by the use of expressions given in this paper is reduced to a rapid and accurate procedure. The methods outlined have been illustrated by the computation by several procedures of the p.e. curve for the $^3\Pi_{0}{}^{+}{}_{u}$ state of the bromine molecule.

§ 1. INTRODUCTION

Several closed expressions have been suggested to facilitate the construction of a potential-energy curve for a particular electronic state of a diatomic molecule (Morse, 1929; Rydberg, 1931; Rosen and Morse, 1932); these have been chosen to lead to an exact solution of the Schrödinger equation (Morse, 1929; Rosen and Morse, 1932) or have been derived as the result of approximation (Rydberg, 1931). For the interpretation of much of the band-spectroscopic data it is necessary to construct more accurate functions than those referred to, and recourse is then had to the laborious graphical procedure of Oldenberg (1929), Rydberg (1931 and 1933) and Klein (1932). This method cannot be applied with accuracy at low vibrational quantum numbers, for which the power-series expansion method of Crawford and Jorgensen (1936) must be employed. The power series for the potential energy in terms of the displacement of the atoms from their equilibrium position does not converge rapidly for any range of displacements, and involves even more computation than the Rydberg-Klein method. Taking advantage of the fact that a cubic in the vibrational quantum number will describe with sufficient accuracy the spectroscopic data for vibrational states, an analytical expression has been derived which is equivalent to the Rydberg-Klein procedure and which facilitates the construction of p.e. curves considerably. In this paper the Rydberg method has been reduced to an analytical operation and the procedure for dealing with cases in which a change in the law of force occurs at high vibrational quantum numbers is outlined.

§2. DERIVATION OF EXPRESSIONS RELATING POTENTIAL ENERGY AND INTERNUCLEAR SEPARATION

The Rydberg-Klein method leads to an expression (Klein, 1932)

$$r_{1,2}(U)=(f/g+f^2)^{1/2}\pm f \qquad \ldots\ldots(1)$$

for the maximum and minimum values $r_{1,2}$ of the interatomic distance for a molecule vibrating with an energy U. f and g are defined in terms of the function

$$S(U,\kappa)=\frac{1}{\pi(2\mu)^{1/2}}\int_0^{I'}\{U-E(I,\kappa)\}^{1/2}\,dI, \qquad \ldots\ldots(2)$$

in which $E(I,\kappa)$ is the sum of the vibrational and rotational energy of the molecule,

$$I=\mathbf{h}(v+\tfrac{1}{2})^{*},$$

$$\kappa=\frac{J(J+1)\mathbf{h}^2}{8\pi^2\mu}=\frac{J(J+1)}{b}{}^{*},$$

v the vibrational quantum number,
J the rotational quantum number,
μ the reduced mass of the molecule,
and $I=I'$ when $U=E$.

f and g are then defined by

$$f=\frac{\partial S}{\partial U},$$

$$g=-\frac{\partial S}{\partial \kappa}.$$

Since $E(I,\kappa)$ can be expressed in terms of v and J and the derived constants ω_e, $\omega_e x_e$, $\omega_e y_e$, B_e, α and D_e to the accuracy of the experimental data, in some cases by a quadratic in v only, but in most by a cubic, then the expression to be integrated is known. If the integration can be performed, the function $r(U)$ can be obtained in terms of the band-spectroscopic constants.

$E(I,\kappa)$ a quadratic in I

$E(I,\kappa)$ may be expressed approximately as

$$\omega_e(v+\tfrac{1}{2})-\omega_e x_e(v+\tfrac{1}{2})^2+B_e J(J+1)+D_e J^2(J+1)^2-\alpha J(J+1)(v+\tfrac{1}{2}). \qquad \ldots\ldots(3)$$

Substituting this into (2) and introducing the variables I and κ, we have

$$S(U,\kappa)=q\int_0^{I'}(A-lI+mI^2)^{1/2}\,dI,$$

where
$$1/q=\pi(2\mu\mathbf{h})^{1/2},$$
$$A=\mathbf{h}\{U-B_e J(J+1)-D_e J^2(J+1)^2\},$$
$$l=\omega_e-\alpha J(J+1),$$
and
$$m=\frac{\omega_e x_e}{\mathbf{h}}.$$

---

[*] The expressions given here are the quantum-mechanical equivalents of the classical quantities I and κ.

Differentiating with respect to U and κ under the integral sign and then performing the integrations leads to the following expressions for f and g:

$$f=\frac{q\boldsymbol{h}}{2m^{1/2}}\log_e\left\{\frac{(l^2-4mA)^{1/2}}{l-(4mA)^{1/2}}\right\},$$

$$g=\frac{qb}{4m^{3/2}}\left[\alpha(4mA)^{1/2}+(2mB_e\boldsymbol{h}+4mD_e\boldsymbol{h}b\kappa-\alpha l)\log_e\left\{\frac{(l^2-4mA)^{1/2}}{l-(4mA)^{1/2}}\right\}\right],$$

which for U and the constants ω_e, $\omega_e x_e$, α, B_e and D_e expressed in wave numbers and for the rotationless state ($J=0$) become

$$f\,(\text{cm.})=\left(\frac{\boldsymbol{h}}{8\pi^2\mathbf{c}\mu\,.\,\omega_e x_e}\right)^{1/2}\log_e\left\{\frac{(\omega_e{}^2-4\omega_e x_e U)^{1/2}}{\omega_e-(4\omega_e x_e U)^{1/2}}\right\}, \qquad \ldots\ldots(4)$$

$$g\,(\text{cm.}^{1})=\left(\frac{2\pi^2\mu\mathbf{c}}{\boldsymbol{h}(\omega_e x_e)^3}\right)^{1/2}\left[\alpha(4\omega_e x_e U)^{1/2}\right.$$
$$\left.+(2\omega_e x_e B_e-\alpha\omega_e)\log_e\left\{\frac{(\omega_e{}^2-4\omega_e x_e U)^{1/2}}{\omega_e-(4\omega_e x_e U)^{1/2}}\right\}\right]. \qquad \ldots\ldots(5)$$

E(I, κ) a cubic in I

If the law of force remains unchanged for all vibrational levels, then an expression of the form

$$\omega_e(v+\tfrac{1}{2})-\omega_e x_e(v+\tfrac{1}{2})^2+\omega_e y_e(v+\tfrac{1}{2})^3+B_eJ(J+1)$$
$$+D_eJ^2(J+1)^2-\alpha J(J+1)(v+\tfrac{1}{2}) \qquad \ldots\ldots(6)$$

will describe the energies of all the vibrational-rotational states adequately. The function $S(U,\kappa)$ then becomes

$$q\int_0^{I'}(A-lI+mI^2+nI^3)^{1/2}\,dI,$$

where

$$n=-\frac{\omega_e y_e}{\boldsymbol{h}^2}$$

and A, l, m and q are defined as before. It is convenient to differentiate $S(U,\kappa)$ with respect to U and κ before integrating with respect to I. We then obtain

$$f=\frac{\partial S}{\partial U}=\frac{q\boldsymbol{h}}{2}\int_0^{I'}\frac{dI}{(A-lI+mI^2+nI^3)^{1/2}} \qquad \ldots\ldots(7)$$

and

$$g=-\frac{\partial S}{\partial\kappa}=\frac{q(B_eb\boldsymbol{h}+2D_eb^2\boldsymbol{h}\kappa)}{2}\int_0^{I'}\frac{dI}{(A-lI+mI^2+nI^3)^{1/2}}$$
$$-\frac{\alpha bq}{2}\int_0^{I'}\frac{I\,.\,dI}{(A-lI+mI^2+nI^3)^{1/2}} \qquad \ldots\ldots(8)$$

(1) *Expressions for f*

By making the substitution $z=I+m/3n$ we obtain

$$f=\frac{q\boldsymbol{h}}{n^{1/2}}\int_{m/3n}^{I'+m/3n}\frac{dz}{(4z^3-g_2z-g_3)^{1/2}},$$

where

$$g_2=4\left(\frac{l}{n}+\frac{m^2}{3n^2}\right)$$

and

$$g_3=-4\left(\frac{lm}{3n^2}+\frac{2}{27}\frac{m^3}{n^3}+\frac{A}{n}\right)$$

are the invariants of the cubic.

Making the substitution $z=\wp(u\,;\,g_2,g_3)$, where $\wp(u\,;\,g_2,g_3)$ is the Weierstrassian elliptic function,

$$f=\frac{qh}{n^{1/2}}\int_{u_1}^{u_2}\frac{\wp'(u)\,du}{(4\wp^3(u)-g_2\wp(u)-g_3)^{1/2}}$$
$$=\frac{qh}{n^{1/2}}(u_2-u_1),$$

since $\wp'^2(u)=4\wp^3(u)-g_2\wp(u)-g_3$.

The form of the solution is determined by (i) the sign of the discriminant $\Delta(=g_2{}^3-27g_3{}^2)$ of the cubic, and (ii) the sign of $\omega_e y_e$.

A real dissociation energy is possible only for the two cases in which $\Delta>0$; however, where discontinuities occur in the law of force, ranges of v-values for which $\Delta<0$ may be encountered; the solutions for these conditions will therefore be indicated. The solutions will be given in detail for $\omega_e y_e$ negative only; those for $\omega_e y_e$ positive are given in table 1.

(*a*) *For* $\omega_e y_e$ *negative*; $\Delta>0$,

$$\wp(u)=e_3+(e_1-e_3)\,\mathrm{ns}^2\,(u(e_1-e_3)^{1/2}\,|\,k),$$

where
$$k=\left(\frac{e_2-e_3}{e_1-e_3}\right)^{1/2}$$

is the modulus of the Jacobian elliptic function ns, and e_1, e_2 and e_3 are the real roots of the cubic, such that $e_1>e_2>e_3$ and $e_1+e_2+e_3=0$.

When $I=I'$, $z=e_2=\wp(u)$. Since $e_2\rightarrow m/3n$ as $U\rightarrow 0$, e_2 is the physically significant root.

Hence
$$u_2=(e_1-e_3)^{-1/2}\,\mathrm{sn}^{-1}\,(1/k\,|\,k),$$
and since k is always <1, u_2 is unreal and may be written
$$u_2=(e_1-e_3)^{-1/2}(K+iK').$$

K and K' are the complete elliptic integrals defined by the hypergeometric functions

$$\frac{1}{2\pi}F(\tfrac{1}{2},\tfrac{1}{2}\,;\,1\,;\,k^2)\quad\text{and}\quad\frac{1}{2\pi}F(\tfrac{1}{2},\tfrac{1}{2}\,;\,1\,;\,k'^2)$$

respectively $(k'^2=1-k^2)$.

When $I=0$, $z=m/3n=\wp(u)$, so that

$$u_1=(e_1-e_2)^{-1/2}\,\mathrm{sn}^{-1}\left\{\left(\frac{e_1-e_3}{m/3n-e_3}\right)^{1/2}\middle|\,k\right\},$$

and since $e_3<0$ and $e_1>m/3n$, u_1 is complex also.

However, using the relation (Whittaker and Watson, 1940c)

$$\mathrm{sn}\,(v+iK'\,|\,k)=k^{-1}\,\mathrm{ns}\,(v\,|\,k),$$

we obtain
$$u_1=(e_1-e_3)^{-1/2}\left[\mathrm{sn}^{-1}\left\{\left(\frac{m/3n-e_3}{e_2-e_3}\right)^{1/2}\middle|\,k\right\}+iK'\right].$$

Expressing U and the spectroscopic constants in wave numbers, we obtain

$$f=q\left(\frac{h}{-\omega_e y_e\,.\,c}\right)^{1/2}(\theta_1-\phi_1),\qquad\ldots\ldots(9)$$
$$\theta_1=(e_1-e_3)^{-1/2}K,$$
$$\phi_1=(e_1-e_3)^{-1/2}\,\mathrm{sn}^{-1}\left\{\left(\frac{m/3n-e_3}{e_2-e_3}\right)^{1/2}\middle|\,k\right\},$$

which is real.

Table 1

$$f = q\left(\frac{h}{-\omega_e y_e \, . \, c}\right)^{1/2}(\theta-\phi),$$

$$g = 4\pi\left(\frac{2\mu c}{-\omega_e y_e}\right)^{1/2}\left[\left(B_e - \frac{\alpha \, . \, \omega_e x_e}{3\omega_e y_e}\right)(\theta-\phi) + \frac{\alpha}{h}(\zeta(\theta)-\zeta(\phi)+\psi)\right]$$

Δ	${}_ey_e$	Real roots	Physically significant root	θ	ϕ	ψ
>0	<0	$e_1>e_2>e_3$	$e_2(>0)$	$K(e_1-e_3)^{-1/2}$	$(e_1-e_3)^{-1/2}\mathrm{sn}^{-1}\left\{\left(\frac{m/3n-e_2}{e_1-e_3}\right)^{\frac{1}{2}}\Big\vert k\right\}$	$\left\{\frac{(e_2-m/3n)(e_1-m/3n)}{(m/3n-e_3)}\right\}^{1/2}$
<0	<0	e_2	$e_2(>0)$	$KH^{-1/2}$	$\frac{1}{2}H^{-1/2}\mathrm{cn}^{-1}\left\{-\left(\frac{H-m/3n+e_2}{H+m/3n-e_2}\right)\Big\vert k\right\}$	0
>0	>0	$e_1>e_2>e_3$	$e_3(<0)$	$iK'(e_1-e_3)^{-1/2}$	$i(e_1-e_3)^{-1/2}\mathrm{sn}^{-1}\left\{\left(\frac{e_1-e_2}{e_1-m/3n}\right)^{\frac{1}{2}}\Big\vert k'\right\}^*$	0
<0	>0	e_2	$e_2(<0)$	$iK'H^{-1/2}$	$\frac{i}{2}H^{-1/2}\mathrm{cn}^{-1}\left\{-\left(\frac{H+m/3n-e_2}{H-m/3n+e_2}\right)\Big\vert k'\right\}^*$	0

* Where θ and ϕ are imaginary quantities, both f and g are real, since $\zeta(iu)=i \, . \, \zeta(u)$.

(*b*) *For $\omega_e y_e$ negative*; $\Delta<0$, there is only one real root, designated e_2, and

$$\wp(u)=e_2+H\frac{1+\mathrm{cn}\,(2uH^{1/2}|k)}{1-\mathrm{cn}\,(2uH^{1/2}|k)},$$

where $$H^2=2e_2{}^2+g_3/4e_2$$

and $$k^2=\tfrac{1}{2}-\tfrac{3}{4}\,.\,e_2/H,$$

from which we obtain (U, ω_e, etc., in cm.$^{-1}$),

$$f=q\left(\frac{h}{-\omega_e y_e\,.\,c}\right)^{1/2}(\theta_2-\phi_2), \qquad \ldots\ldots(10)$$

$$\theta_2=H^{-1/2}K,$$

$$\phi_2=\tfrac{1}{2}H^{-1/2}\,\mathrm{cn}^{-1}\left\{-\left(\frac{H-m/3n+e_2}{H+m/3n-e_2}\right)\middle|\,k\right\}.$$

(2) *Expressions for g*

Inspection of equation (8) shows that the integral involved in the first term of g is identical with that for f, and it may be written, for $\Delta>0$ and $\omega_e y_e$ negative, as

$$\frac{qbh(B_e-2D_e b\kappa)}{n^{1/2}}(\theta_1-\phi_1).$$

On making the substitution $I+m/3n=z=\wp(u;\ g_2,g_3)$, the second term becomes

$$\frac{\alpha bq}{n^{1/2}}\cdot\frac{m}{3n}(\theta_1-\phi_1)-\frac{\alpha bq}{n^{1/2}}\int_{u_1}^{u_2}\wp(u)\,.\,du;$$

$$\int_{u_1}^{u_2}\wp(u)\,.\,du=\zeta(u_2)-\zeta(u_1),$$

where $\zeta(u)$ is the Weierstrassian ζ-function.

Although u_1 and u_2 are complex variables, by making use of the relations

$$\zeta(u+v)=\zeta(u)+\zeta(v)+\tfrac{1}{2}\left\{\frac{\wp'(u)-\wp'(v)}{\wp(u)-\wp(v)}\right\}$$

and $$\wp(u;\ g_2,g_3)=\lambda\,.\,\wp(u\lambda^{1/2};\ g_2\lambda^{-2},\ g_3\lambda^{-3})$$

(Whittaker and Watson, 1940a and b), the pure imaginary parts of the functions may be separated and cancelled, leading to

$$\zeta(u_1)-\zeta(u_2)=\zeta(\theta_1)-\zeta(\phi_1)+\psi;\quad\left\{\frac{(e_2-m/3n)(e_1-m/3n)}{(m/3n-e_3)}\right\}^{1/2}=\psi,$$

which is real.

For the rotationless state ($\kappa=0$), and with U, ω_e, etc. in cm.$^{-1}$, we have

$$g=4\pi\left(\frac{2\mu c}{-\omega_e y_e}\right)^{1/2}\left[\left(B_e-\frac{\alpha\omega_e x_e}{3\omega_e y_e}\right)(\theta_1-\phi_1)+\frac{\alpha}{h}(\zeta(\theta_1)-\zeta(\phi_1)+\psi)\right] \qquad \ldots\ldots(11)$$

For a negative discriminant ($\Delta<0$), and $\kappa=0$,

$$g=4\pi\left(\frac{2\mu c}{-\omega_e y_e}\right)^{1/2}\left[\left(B_e-\frac{\alpha\omega_e x_e}{3\omega_e y_e}\right)(\theta_2-\phi_2)+\frac{\alpha}{h}(\zeta(\theta_2)-\zeta(\phi_2))\right] \qquad \ldots\ldots(12)$$

The solutions for $\omega_e y_e$ positive are given in table 1.

§3. THE DISSOCIATION ENERGY

The condition for $U=\boldsymbol{D}_e$*, the dissociation energy, is that f, the half-width of the $U(r)$ curve at an energy U, should become infinite. For the quadratic case, this can be seen to occur when $\omega_e{}^2=4\omega_e x_e\,.\,U$ in equation (4), i.e.

$$\boldsymbol{D}_e=\frac{\omega_e{}^2}{4\omega_e x_e},$$

as expected.

For the cubic case, only $\Delta>0$ can lead to a real dissociation energy. In equation (9), $f\to\infty$ as $K\to\infty$, which occurs when $k\to1$. Since

$$k=\left(\frac{e_2-e_3}{e_1-e_3}\right)^{1/2},$$

$k=1$ only when $e_1=e_2$, i.e. when $\Delta=0$ or $27g_3{}^2=g_2{}^3$.

We then have

$$U=\boldsymbol{D}_e=\frac{1}{27\omega_e y_e{}^2}\{2(\omega_e x_e{}^2-3\omega_e\,.\,\omega_e y_e)^{3/2}-\omega_e x_e(2\omega_e x_e{}^2-9\omega_e\,.\,\omega_e y_e)\}, \qquad \ldots\ldots(13)$$

which is consistent with the value obtained by equating $\partial E/\partial v$ to zero to obtain $v_{\max}$ and substituting to obtain $E_{v_{\max}}=\boldsymbol{D}_e$.

§4. THE EQUILIBRIUM INTERNUCLEAR SEPARATION

From equation (1) it is evident that the equilibrium internuclear distance r_e corresponds to the limit

$$\operatorname*{Lt}_{U\to0}\{(f/g+f^2)^{1/2}\pm f\}=\frac{1}{\operatorname*{Lt}_{U\to0} g/f}.$$

Equations (4) and (5) are readily shown to lead to

$$r_e=\left(\frac{h}{8\pi^2\mathbf{c}B_e\mu}\right)^{1/2},$$

in agreement with definition.

For the cubic case, equations (9) and (11) lead to

$$\operatorname*{Lt}_{U\to0} g/f=b\left(B_e+\frac{\alpha m}{3n}\right)-\frac{\alpha b}{h}\,.\,e_1+\frac{\alpha b}{h}(e_1-m/3n)$$
$$=b\,.\,B_e.$$

As $b=\dfrac{8\pi^2\mu}{h^2}$, $r_e=\left(\dfrac{h}{8\pi^2\mathbf{c}B_e\mu}\right)^{1/2}$, as before.

§5. THE MORSE FUNCTION

The function proposed by Morse may be shown to lead to the same expression for f as equation (4). The Morse function may be written as

$$(r_{1,2}-r_e)=-\frac{1}{a}\log_e\left\{1\pm\left(\frac{U}{\boldsymbol{D}_e}\right)^{1/2}\right\},$$

where
$$a=2\pi\left(\frac{2\mu\,.\,\omega_e x_e\,.\,\mathbf{c}}{h}\right)^{1/2}.$$

---

* $\boldsymbol{D}_e$, the dissociation energy, should not be confused with D_e, one of the coefficients in the rotational term.

The width of this curve at a vibrational level of potential energy U is

$$2f=(r_2-r_e)-(r_1-r_e)$$
$$=\frac{1}{a}\log_e\left\{\frac{\boldsymbol{D}_e-U}{(\boldsymbol{D}_e^{1/2}-U^{1/2})^2}\right\},$$

whence
$$f=\left(\frac{\boldsymbol{h}}{8\pi^2\boldsymbol{c}\mu\omega_e x_e}\right)^{1/2}\log_e\left\{\frac{(\boldsymbol{D}_e-U)^{1/2}}{\boldsymbol{D}_e^{1/2}-U^{1/2}}\right\},$$

which is identical with equation (4). This equivalence is to be expected, since the Morse function leads to an exact solution of the one-dimensional Schrödinger equation and describes the energy levels E_v as a quadratic in $(v+\frac{1}{2})$, whereas in the derivation of equation (4) this has been used as the fundamental assumption.

§6. DISCONTINUITY IN THE LAW OF FORCE

Birge (1929b) has pointed out that a discontinuity in the law of force for a certain molecular vibrational energy is of frequent occurrence and, on either side of this discontinuity, the use of functions having different values of ω_e, $\omega_e x_e$, etc. is necessitated if the band spectroscopic data are to be satisfied. In the general case, then, we have

$$0<I<I_1;\qquad A_1, l_1, m_1, n_1;$$
$$I_1<I<I_2;\qquad A_2, l_2, m_2, n_2.$$
$$\cdots\cdots\cdots\cdots$$

Writing $(A_i-l_iI+m_iI^2+n_iI^3)^{-1/2}=\chi_i$,

$$f=\frac{q\boldsymbol{h}}{2}\left\{\int_{I_r}^{I'}\chi_{r+1}\,.\,dI+\sum_{i=1}^{i=r}\int_{I_{i-1}}^{I_i}\chi_i\,.\,dI\right\},\qquad\ldots\ldots(15)$$

$$g=\frac{q(B_eb\boldsymbol{h}+2D_eb^2\boldsymbol{h}\kappa)}{2}\left\{\int_{I_r}^{I'}\chi_{r+1}\,.\,dI+\sum_{i=1}^{i=r}\int_{I_{i-1}}^{I_i}\chi_i\,.\,dI\right\}$$
$$-\frac{q\alpha b}{2}\left\{\int_{I_r}^{I'}\chi_{r+1}\,.\,I\,.\,dI+\sum_{i=1}^{i=r}\int_{I_{i-1}}^{I_i}\chi_i\,.\,I\,.\,dI\right\}\qquad\ldots\ldots(16)$$

in which $I_0=0$.

§7. THE CONSTRUCTION OF P.E. CURVES

It is apparent that the potential function $U(r)$ can now be constructed from the spectroscopic constants by means of the expressions (9) and (11), (10) and (12) or (15) and (16) as required. The Jacobian elliptic functions sn and cn and the elliptic integral K have been tabulated (Milne-Thomson, 1931) and a double inverse interpolation of these tables is involved in the computation; the Weierstrassian ζ-function, however, has been tabulated for the equianharmonic case only and it must therefore be evaluated by means of the series

$$\zeta(u)=\frac{1}{u}-\frac{g_2u^3}{60}-\frac{g_3u^5}{140}-\frac{g_2{}^2u^7}{8400}-\frac{g_2g_3}{18480}u^9$$
$$-\left(\frac{g_2{}^3}{1\cdot716\times10^6}+\frac{g_3{}^2}{1\cdot12112\times10^5}\right)u^{11}-\frac{g_2{}^2g_3}{2\cdot4024\times10^6}u^{13}-\ldots\qquad\ldots\ldots(17)$$

Normally, evaluation to the term in u^{13} is sufficient. This computation can be performed more rapidly and accurately than the Rydberg graphical procedure, although the evaluation of the ζ-function is laborious.

A rapid method, with somewhat less theoretical justification, but with adequate accuracy, may be used to avoid the evaluation of the ζ-function. Since $\omega_e y_e$ is invariably much smaller than $\omega_e x_e$, it is possible to represent the band data by a small number of quadratic expressions in $(v+\frac{1}{2})$ satisfactorily. The method is then analogous to that outlined in § 6, but some of the points of discontinuity in the apparent ω_e and $\omega_e x_e$ values will be artificial and will not necessarily indicate a change in the law of force. After fixing the points of discontinuity by inspection of a plot of $(G(v))/(v+\frac{1}{2})$ against $(v+\frac{1}{2})$, the constants ω_e and $\omega_e x_e$ for each continuous range may be evaluated by the least squares method. The subsequent calculation of $r_{1,2}(U)$ is extremely rapid.

§ 8. ILLUSTRATIVE CALCULATION

Using Brown's (1931) data for the ${}^3\Pi_{0^+u}$ state of ${}^{79}\mathrm{Br}{}^{81}\mathrm{Br}$, the potential-energy curve for the rotationless state ($J=0$) has been evaluated up to the level $v'=21$ by several different procedures for comparison. The procedures employed were:—

(i) Using the values $\omega_e = 165{\cdot}4$ cm.$^{-1}$, $\omega_e x_e = 1{\cdot}577$ cm.$^{-1}$, $\omega_e y_e = -0{\cdot}0087$ cm.$^{-1}$, $B_e = 0{\cdot}0595$ cm.$^{-1}$ and $\alpha = 6{\cdot}25\times 10^{-4}$ cm.$^{-1}$ in equation (6) to express the spectroscopic data adequately over the range $v'=0$ to $v'=21$, the quantities f and g were evaluated by means of equations (9) and (11). r_1 and r_2 values were then calculated by means of equation (1) for each $U=E_{v'}$.

(ii) The values $\omega_e = 165{\cdot}4$ cm.$^{-1}$, $\omega_e x_e = 1{\cdot}59$ cm.$^{-1}$, $r_e = 2{\cdot}656\times 10^{-8}$ cm. and the extrapolated value of $D_e = 3814$ cm.$^{-1}$ were used to construct a Morse curve.

(iii) Using $\omega_e = 165{\cdot}4$ cm.$^{-1}$ and $\omega_e x_e = 1{\cdot}59$ cm.$^{-1}$, the potential-energy curve was computed from equations (4) and (5). This procedure is equivalent to constructing a Morse curve in which

$$D_e = \frac{\omega_e^2}{4\omega_e x_e}.$$

(iv) Using values of ω_e, $\omega_e x_e$ and derived D_e and r_e values appropriate to each pair of levels $E_{v'}$ and $E_{v'+1}$, separate Morse functions have been evaluated for each region $E_{v'} < U < E_{v'+1}$, so that the complete potential-energy curve is built up from 21 segments of separate Morse curves. This is equivalent to assuming an artificial discontinuity in the law of force at each level and calculating according to equations (15) and (16) in which $\chi_i = (A_i - l_i I + m_i I^2)^{-1/2}$.

The results obtained by application of each of these methods are tabulated for $U=E_{v'}$ up to $v'=21$ in table 2 and plotted in the figure for comparison.

It is evident that method (i) and (iv) give results within the accuracy expected from method (i), which is limited by the incomplete evaluation of the Weierstrassian ζ-function (taken to the term in u^{17}), and that both the Morse functions (ii) and (iii) diverge markedly from the correct function, particularly on the outer limit and at the higher vibrational levels.

Brown (1931) found a discontinuity in the law of force to occur at $v'=21$; examination of his data indicates that a second discontinuity occurs at $v'=37$.

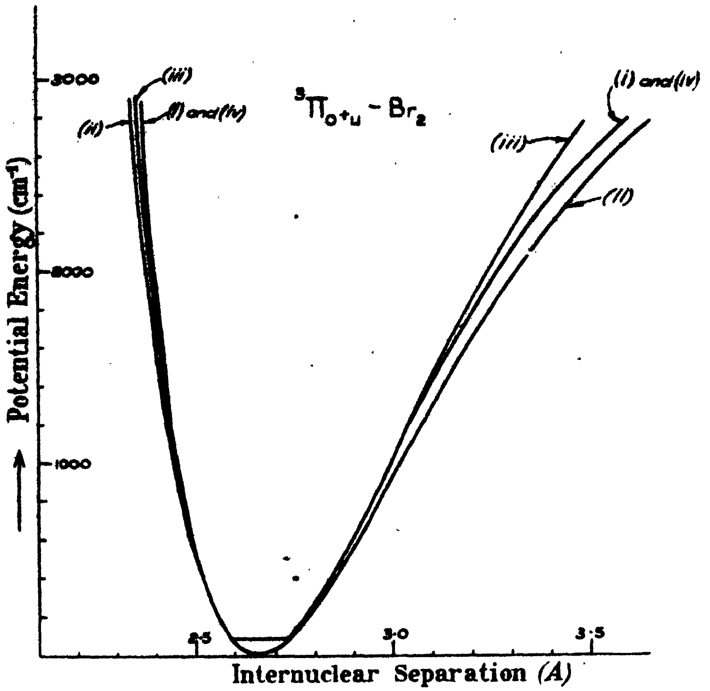


Table 2

v'	U (cm.⁻¹)	r_1 (A.)				r_2 (A.)			
		(i)	(ii)	(iii)	(iv)	(i)	(ii)	(iii)	(iv)
0	82·3	2·589	2·586	2·589	2·590	2·731	2·738	2·733	2·732
1	244·5	2·548	2·547	2·549	2·550	2·797	2·805	2·796	2·796
2	403·5	2·516	2·511	2·519	2·519	2·844	2·858	2·844	2·845
3	559·2	2·495	2·489	2·498	2·499	2·884	2·903	2·888	2·887
4	711·6	2·478	2·471	2·481	2·482	2·922	2·947	2·924	2·925
5	860·5	2·464	2·456	2·467	2·467	2·959	2·975	2·960	2·962
6	1006·1	2·452	2·443	2·454	2·455	2·997	3·027	2·997	3·000
7	1148·1	2·442	2·431	2·442	2·444	3·028	3·066	3·030	3·034
8	1286·7	2·433	2·420	2·432	2·435	3·065	3·104	3·063	3·069
9	1421·5	2·426	2·411	2·420	2·426	3·096	3·141	3·094	3·101
10	1552·7	2·419	2·403	2·412	2·418	3·134	3·178	3·126	3·138
11	1680·3	2·413	2·395	2·404	2·411	3·171	3·216	3·160	3·176
12	1804·0	2·408	2·387	2·399	2·405	3·208	3·255	3·192	3·214
13	1924·0	2·402	2·381	2·390	2·399	3·242	3·295	3·225	3·250
14	2040·2	2·398	2·374	2·384	2·394	3·282	3·334	3·260	3·288
15	2152·3	2·395	2·368	2·380	2·389	3·325	3·373	3·288	3·323
16	2260·6	2·389	2·362	2·376	2·384	3·360	3·413	3·320	3·361
17	2364·9	2·385	2·358	2·372	2·380	3·401	3·452	3·350	3·399
18	2465·0	2·383	2·354	2·367	2·376	3·444	3·493	3·381	3·438
19	2561·1	2·379	2·349	2·363	2·373	3·486	3·538	3·414	3·479
20	2652·9	2·376	2·344	2·358	2·370	3·532	3·581	3·447	3·531
21	2740·6	2·374	2·340	2·355	2·368	3·575	3·627	3·477	3·579

h and *c* values throughout taken from Birge (1929 *a*), so that data are comparable with those published previously.

The values of ω_e, $\omega_e x_e$ and $\omega_e y_e$ appropriate to levels below and above $v'=37$ were evaluated by the method of least squares to give

$$21 \leqslant v' \leqslant 37 \qquad G(v') = 172{\cdot}70(v'+\tfrac{1}{2}) - 2{\cdot}212(v'+\tfrac{1}{2})^2 + 4{\cdot}98 \times 10^{-3}(v'+\tfrac{1}{2})^3,$$

$$37 \leqslant v' \leqslant 48 \qquad G(v') = 184{\cdot}64(v'+\tfrac{1}{2}) - 2{\cdot}857(v'+\tfrac{1}{2})^2 + 1{\cdot}373 \times 10^{-2}(v'+\tfrac{1}{2})^3.$$

If we assume that no further discontinuity occurs beyond $v'=37$, then we may use the values of ω_e, $\omega_e x_e$ and $\omega_e y_e$ for $v'>37$ to compute the dissociation energy by means of equation (13). The value obtained, 3808 cm^{-1}, is in close agreement with the value 3814 cm^{-1} obtained by Brown using the graphical extrapolation method of Birge. The calculated value of $v'_{\max}$ is 51.

§ 9. ACKNOWLEDGMENTS

I am indebted to Messrs. R. H. Dalitz and A. K. Head for many helpfu suggestions.

REFERENCES

Birge, R. T., 1929 a. *Phys. Rev.*, Suppl. (*Rev. Mod. Phys.*), **1**, 1.
Birge, R. T., 1929 b. *Trans. Faraday Soc.*, **25**, 707.
Brown, W. G., 1931. *Phys. Rev.*, **38**, 1179.
Crawford, F. H. and Jorgensen, T., 1936. *Phys. Rev.*, **49**, 745.
Klein, O., 1932. *Z. Phys.*, **76**, 226.
Milne-Thomson, L. M., 1931. *Die elliptischen Funktionen von Jacobi* (Berlin).
Morse, P. M., 1929. *Phys. Rev.*, **34**, 57.
Oldenberg, O., 1929. *Z. Phys.*, **56**, 563.
Rosen, N. and Morse, P. M., 1932. *Phys. Rev.*, **42**, 210.
Rydberg, R., 1931. *Z. Phys.*, **73**, 376.
Rydberg, R., 1933. *Z. Phys.*, **80**, 514.
Whittaker, E. T. and Watson, G. N., 1940 a. *Modern Analysis* (Cambridge), p. 439 ; 1940 b. *Ibid.*, p. 451 ; 1940 c. *Ibid.*, p. 503.

NOTE ON THE INTERPRETATION OF THE VISIBLE ABSORPTION SPECTRUM OF BROMINE

By A. L. G. REES,
Division of Industrial Chemistry, Council of Scientific and Industrial Research, Melbourne, Australia

MS. received 15 *January* 1947

ABSTRACT. The accurate construction of the lower parts of the potential-energy curves for the excited states involved in the transitions responsible for the visible absorption spectrum of bromine has enabled a choice to be made between two possible alternative interpretations.

The interpretation of the long-wavelength spectra of the halogen molecules has been discussed in detail by Mulliken (1940), who has found it possible to assign transitions to the various observed continua with certainty in all cases but bromine, where existing experimental and theoretical data do not allow a decision between two alternatives to be made. The possible interpretations of the two overlapping continua in the visible absorption of bromine vapour, designated A ($\nu_{max} = 24\,300\ \text{cm}^{-1}$) and B ($\nu_{max} = 20\,740\ \text{cm}^{-1}$) (Acton, Aickin and Bayliss, 1936), are:—

(i) A $^1\Sigma_g^+ \rightarrow {}'\Pi_u$,
B $^1\Sigma_g^+ \rightarrow {}^3\Pi_{0^+u}$;
$^1\Sigma_g^+ \rightarrow {}^3\Pi_{1u}$ calculated to have $\nu_{max} = 18\,630\ \text{cm}^{-1}$ and to be relatively very weak.

(ii) A $^1\Sigma_g^+ \rightarrow {}^1\Pi_u$,
B $^1\Sigma_g^+ \rightarrow {}^3\Pi_{1u}$;
$^1\Sigma_g^+ \rightarrow {}^3\Pi_{0^+u}$ calculated to occur at $l\nu_{max} = 22\,800\ \text{cm}^{-1}$ and to be approximately equal to intensity in $^1\Sigma_g^+ \rightarrow {}^3\Pi_{1u}$.

Franck-Condon evidence (Bayliss, 1937; Darbyshire, 1937) appears to be in favour of interpretation (i). The evidence is not conclusive, since the lower part of the $^3\Pi_{0^+u}$ curve, fitted by Bayliss to the upper segment of the curve obtained from the B continuum, was constructed by the use of the Morse function, which deviates from the true curve at high vibrational levels. Moreover, the weak continuum arising from the $^1\Sigma_g^+ \rightarrow {}^3\Pi_{1u}$ transition is expected in the region of the band absorption and has not been detected experimentally. On the other hand, certain of the effects observed in bromine solutions (Aickin, Bayliss and Rees, 1938) could be interpreted in terms of (ii) rather than (i) (Bayliss and Rees, 1939). As Mulliken (1940) has pointed out, an unambiguous decision could be made by an accurate construction of the lower parts of the $^3\Pi_{0^+u}$ and $^3\Pi_{1u}$ curves from the existing band-spectroscopic data. Accurate potential-energy curves may now be constructed by analytical methods described in the preceding paper (Rees, 1947), and the laborious graphical procedure of the Rydberg-Klein method avoided.

As there were some points of disagreement in the independent evaluation of the spectroscopic constants from the band data by Brown (1931) and Darbyshire (1937) respectively, the derivation of the constants has been re-examined. For $^3\Pi_{0^+u}$, a least-squares calculation shows that a cubic in $(v' + \frac{1}{2})$, with numerical values of the coefficients as given by Brown, is necessary to describe the data up to $v' = 21$, although Darbyshire claims that a quadratic is adequate up to $v' = 15$. Darbyshire's value of E_e for this state appears to be in error by the quantity $2G'(0)$; the correct value is $15\,749\ \text{cm}^{-1}$ For $^3\Pi_{1u}$ the data lead to a value for E_e of $13\,856\ \text{cm}^{-1}$ and not $13\,814\ \text{cm}^{-1}$ as quoted by Darbyshire. These values are important in fixing the $^3\Pi_{1u} - {}^3\Pi_{0^+u}$ interval (now $1893\ \text{cm}^{-1}$ and *not* $2104\ \text{cm}^{-1}$) and the relationship of these states to the ground state potential-energy curve.

Using these data for $^{79}\text{Br}\,^{81}\text{Br}$, the curve for $^3\Pi_{0^+u}$ was constrcuted up to $v' = 22$ (within $170\ \text{cm}^{-1}$ of D_e' for this state) and for $^3\Pi_{1u}$ up to $v' = 22$ (within $150\ \text{cm}^{-1}$ of D' for this state) by method (iv) of the preceding paper (Rees, 1947). For $^3\Pi_{0^+u}$ the curve up to $v' = 21$ was calculated also by method (i), employing the analytical expressions equivalent to the Rydberg-Klein graphical method. These curves are plotted in the figure, together with the upper segments derived by Bayliss (1937) from the analysis of the continuous absorption data. The lower part of the $^3\Pi_{0^+u}$ curve extrapolates perfectly (within 0·005 A.) on to Bayliss's upper segment corresponding to the B continuum, whereas the corresponding Morse curve is 0·03 A. in error. A reasonable extrapolation of the $^3\Pi_{1u}$ curve would lie about 1900 cm^{-1} below this curve at r_e'' (= 2·284 A.) and could not be extrapolated on to the segment labelled B. The curve for the repulsive $^1\Pi_u$ state has been drawn to cross $^3\Pi_{0^+u}$ in the region of the 3rd and 4th vibrational levels, as suggested previously by Bayliss and Rees (1939).

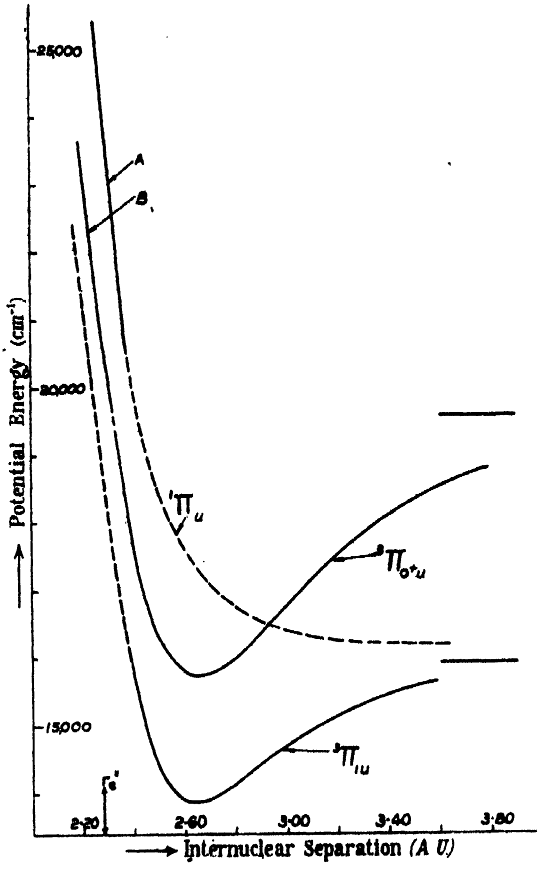


Figure 1. Potential energy curves for the upper states involved in the visible absorption spectrum of Br_2.

This evidence is consistent only with interpretation (i). Moreover, it directly lends strong support to the Franck-Condon principle, as it seems very improbable that the continuity of $U'(r)$ and its first derivative at D_e' should be fortuitous.

REFERENCES

Acton, A. P., Aickin, R. G. and Bayliss, N. S., 1936. *J. Chem. Phys.*, **4**, 474.
Aickin, R. G., Bayliss, N. S. and Rees, A. L. G., 1938. *Proc. Roy. Soc.*, A, **169**, 234.
Bayliss, N. S., 1937. *Proc. Roy. Soc.* A, **158**, 551.
Bayliss, N. S. and Rees, A. L. G., 1939. *J. Chem. Phys.*, **7**, 854.
Brown, W. G., 1931. *Phys. Rev.*, **38**, 1179.
Darbyshire, O., 1937. *Proc. Roy. Soc.*, A, **159**, 93.
Mulliken, R. S., 1940. *Phys. Rev.*, **57**, 500.
Rees, A. L. G., 1947. *Proc. Phys. Soc.*, **59**, 998.

THE EFFECTIVE PERMITTIVITY OF TWO-PHASE SYSTEMS

By D. F. RUSHMAN and M. A. STRIVENS,
Material Research Laboratories, Philips Electrical Ltd., Mitcham

MS. received 7 *February* 1947

ABSTRACT. The effective permittivity of samples of barium titanate with porosities up to 40% v/v. has been determined, and an explanation given in terms of Wiener's mixture law. It is shown that the values of the form factor in this law can only be predicted with certainty over the ranges 0–30% and 70–100% porosity, and takes up intermediate values in the 30–70% porosity region. The observed values of ϵ are in accordance with this hypothesis.

§ 1. INTRODUCTION

It is often required to find the true permittivity of a substance when it is a component of a heterogeneous mixture whose effective permittivity can be measured and where the permittivities of the other components are known. It is also useful for practical applications to be able to calculate the effective permittivity from a knowledge of the volume fractions and the permittivities of the components.

This problem, which also occurs in the magnetic case of the effective permeability of iron-dust cores, etc., has recently been brought into prominence owing to the study of the ferroelectric behaviour of barium titanate, which has only been prepared so far in the form of sintered polycrystalline masses, which are invariably porous to a certain extent.

§ 2. EXPERIMENTAL

Barium metatitanate, $BaTiO_3$, was prepared by milling together equimolecular proportions of the purest available barium carbonate and titanium dioxide, pressing the mixture into blocks and prefiring it at 1250° c. The product was then crushed and re-milled. A chemical analysis at this stage showed the total amount of impurities to be less than 0·2%, with 49·12 mol % BaO and 50·86 mol % TiO_2.

Using this material samples were made by pressing the powder in a die to give discs about 1 cm. × 3 mm. and firing at temperatures ranging between 1000° c. and 1350° c.

This resulted in a series of samples having different porosities and hence different permittivities. The porosity was determined in each case from the expression

$$P = \frac{\rho_x - \rho_A}{\rho_x},$$

where $\rho_x = 6{\cdot}08$ g/cc. is the x-ray density and ρ_A is the apparent (or bulk) density. The latter was determined directly by grinding the samples to a regular shape, weighing them, and measuring them with a micrometer.

Electrodes were applied by evaporation of silver *in vacuo* on to the faces of the discs, and the capacities were measured *in vacuo* at 1·6 Mc/sec. and 20° c. by a substitution method.

The results given in table 1 are shown graphically in figure 1.

Table 1

Sample No.	Firing temp. (° c.)	Density (g/cc.)	Porosity vol. (%)	ϵ_m
1	1130	3·56	41·2	458
2		3·59	40·7	443
3		3·52	41·8	439
4	1150	3·61	40·3	491
5		3·60	40·5	520
6		3·61	40·3	473
7		3·67	39·3	508
8		3·64	39·8	486
9	1200	4·08	32·6	894
10		4·06	32·9	918
11		4·04	33·2	897
12	1250	4·29	29·0	956
13		4·26	29·6	930
14		4·18	31·0	956
15		4·25	29·8	922
16		4·28	29·2	957
17		4·12	31·9	956
18	1300	4·92	18·7	1176
19		4·70	22·3	1160
20		4·68	22·6	1084
21	1350	5·37	11·2	1366
22		5·41	10·5	1381
23		5·45	9·9	1408
24		5·43	10·3	1423
25		5·55	8·3	1447
26		5·45	9·9	1418

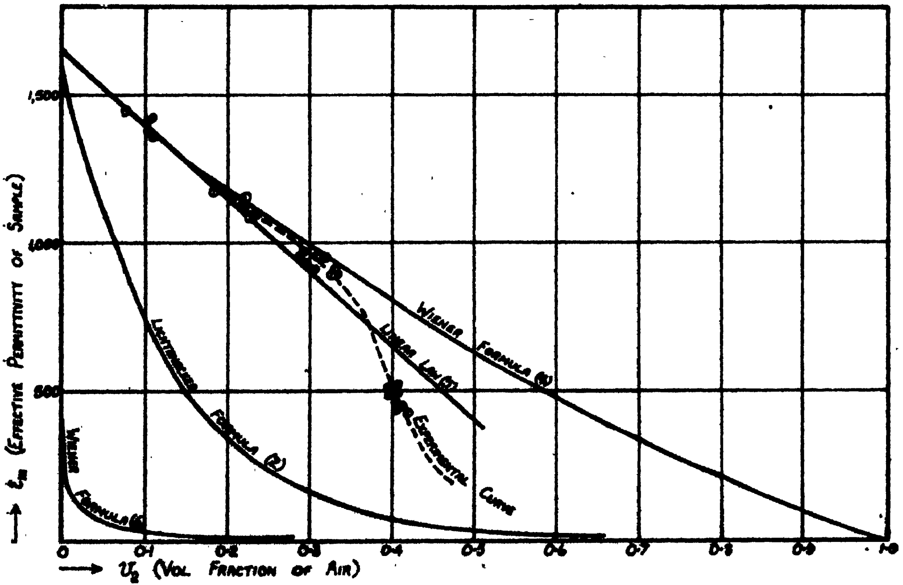


Figure 1.

§ 3. DISCUSSION

Numerous mixture laws are recorded in the literature, some of which have an empirical or semi-empirical basis, others, however, having a more theoretical significance. It is evident from the large quantity of recorded experimental data, and on general grounds also, that several factors, apart from the relative volumes and the permittivities of the constituents, must be accounted for in a general mixture law. The chief of these is the shape or form of the different phases as distinct from their internal structure, which is assumed to be continuous and homogenous and, in particular, which are the continuous and which are the discontinuous phases.

For polarizable spheres in a vacuum the Clausius-Mosotti formula has a sound theoretical basis and leads to the mixture law

$$\frac{\epsilon_m - 1}{\epsilon_m + 2} = \sum_i V_i \frac{\epsilon_i - 1}{\epsilon_i + 2},$$

where V_i is the volume fraction of the constituent of permittivity ϵ_i, and ϵ_m is the effective permittivity of the mixture.

This formula has the support of considerable experimental data, in particular relating to non-polar gases and certain liquids, and is quite valid provided that the concentration of spheres is not too high and that they have neither a random distribution or a cubic symmetry.

The Clausius-Mosotti formula may be generalized to include the case of polarizable spheres embedded in a medium of permittivity ϵ_0. The factor $(\epsilon - 1)/(\epsilon + 2)$ then becomes $(\epsilon - 1)/(\epsilon + 2\epsilon_0)$, and the formula will still be valid within the limits imposed by the model.

Wiener (1904 and 1910) has generalized the formula further and states it in the following form:

$$\frac{\epsilon_m - 1}{\epsilon_m + u} = \sum_i V_i \frac{\epsilon_i - 1}{\epsilon_i + u}, \qquad \ldots\ldots(1)$$

where u is a form factor governed by the shape of the phase present. Table 2 shows the values of u for certain 2-phase systems together with the form which equation (1) assumes. It will be noted that $u = 2\epsilon_0$ for spheres embedded in a medium of permittivity ϵ_0.

A commonly used law is that of Lichtenecker (1926), viz.,

$$\log \epsilon_m = \sum_i V_i \log \epsilon_i, \qquad \ldots\ldots(2)$$

which is derived as an intermediate form of the series and parallel combination laws for dielectrics. Although its theoretical significance is slight, it has met with some success, but only in cases where the difference between the permittivities of the constituents have been small.

In our case $\epsilon_1 \gg \epsilon_2$, where the subscripts 1 and 2 refer to $BaTiO_3$ and air respectively, and reference to figure 1 shows that the observed values show no agreement whatsoever with the Lichtenecker formula.

The observed permittivity decreases almost linearly with increasing porosity until a value of about 35% porosity is reached, after which it falls rapidly. This rather unexpected behaviour is best understood in terms of the Wiener formula (1).

If we regard the pores in barium titanate as spherical (to a first approximation), we may put $u = 2\epsilon_1$, where ϵ_1 is the true permittivity of $BaTiO_3$.

Table 2

Arrangement	*u*-factor	Formula
Layers perpendicular to the lines of force	0	$\frac{1}{\epsilon_m}=\frac{V_1}{\epsilon_1}+\frac{V_2}{\epsilon_2}$
Cylinders of I in II to lines of force	ϵ_2	$\frac{\epsilon_m-\epsilon_2}{\epsilon_m+\epsilon_2}=V_1\frac{\epsilon_1-\epsilon_2}{\epsilon_1+\epsilon_2}$
Spheres of I in II	$2\epsilon_2$	$\frac{\epsilon_m-\epsilon_2}{\epsilon_m+2\epsilon_2}=V_1\frac{\epsilon_1-\epsilon_2}{\epsilon_1+2\epsilon_2}$
Cylinders of II in I to lines of force	ϵ_1	$\frac{\epsilon_m-\epsilon_1}{\epsilon_m+\epsilon_1}=V_2\frac{\epsilon_2-\epsilon_1}{\epsilon_2+\epsilon_1}$
Spheres of II in I	$2\epsilon_1$	$\frac{\epsilon_m-\epsilon_1}{\epsilon_m+2\epsilon_1}=V_2\frac{\epsilon_2-\epsilon_1}{\epsilon_2+2\epsilon_1}$
Layers or discs parallel to lines of force	∞	$\epsilon_m=V_1\epsilon_1+V_2\epsilon_2$

Formula (1), written for a two-phase system, now takes the form

$$\frac{\epsilon_1-\epsilon_m}{\epsilon_m+2\epsilon_1}=V_2\frac{\epsilon_1-\epsilon_2}{\epsilon_2+2\epsilon_1}, \qquad \ldots\ldots(3)$$

where $V_2=(1-V_1)$ is the volume fraction of pores.

In this form Wiener's law should be applicable to any system of spheres dispersed in a continuous medium, but in our case it can be simplified since $\epsilon_1 \gg \epsilon_2$. We have, therefore,

$$\frac{\epsilon_1-\epsilon_m}{\epsilon_m+2\epsilon_1}\simeq\frac{V_2}{2},$$

or, that is,

$$\epsilon_m\simeq\frac{2\epsilon_1(1-V_2)}{(2+V_2)}=\epsilon_1\left[1+3\sum_{n=1}^{\infty}\left(-\frac{V_2}{2}\right)^n\right]. \qquad \ldots\ldots(4)$$

In cases where the total porosity is small, the power series may be taken to only the first term with considerable accuracy, i. e.

$$\epsilon_m\simeq\epsilon_1(1-\tfrac{3}{2}V_2). \qquad \ldots\ldots(5)$$

This approximate form, which has been used elsewhere by us to correct for porosity in high permittivity materials (Rushman and Strivens 1946), has also been derived as a limiting case by Polder and van Santen (1946). Figure 1 shows that the experimental points give good agreement with equation (4) for values of V_2 less than 0·35, and that the approximation (5) is quite valid for extrapolation to zero porosity in this region. We have, however, used equation (4) for our extrapolation and found the true permittivity of this sample of $BaTiO_3$ to be 1650.

The rapid drop in the observed value of ϵ_m in the region of 35% porosity is due to several interdependent factors.

Firstly, there is a limit to the packing density of the air pores. Assuming that they are equal in size and arranged in cubic close-packed symmetry, the limiting porosity beyond which they coalesce is equal to $\pi/3\sqrt{2}$ or about 70%. Below this limiting porosity the Wiener formula (4) with $u=2\epsilon_1$ would apply if the concentration of spheres were not so high that the field in the not too close proximity of each pore ceases to be equivalent to the average or smoothed-out field, which is one of the basic assumptions behind the Clausiu-Mosotti formula. One would expect for this reason alone a breakdown in the Wiener formula (4).

But secondly, it is extremely probable that the air pores have a random distribution and are unequal in size, so that they will begin to coalesce long before the limiting porosity for cubic close-packing, yielding pores shaped like simple or branched filaments, for which the form factor, *u*, cannot be calculated for the general case.

A similar picture holds if we reverse the phases and consider particles of barium titanate dispersed in air. The relevant mixture law is obtained by interchanging subscripts in equation (3), giving

$$\frac{\epsilon_2-\epsilon_m}{\epsilon_m+2\epsilon_2}=V_1\frac{\epsilon_2-\epsilon_1}{\epsilon_1+2\epsilon_2},$$

which for $\epsilon_2=1$ and $\epsilon_1 \gg \epsilon_2$ reduces to

$$\epsilon_m \simeq \epsilon_1\frac{(3-2V_2)}{3+V_2\epsilon_1}. \qquad \ldots\ldots(6)$$

As before, a limiting packing fraction ($V_1 \simeq 0{\cdot}70$) exists beyond which the barium titanate cannot be completely discontinuous. In the intermediate region extending approximately from 30% to 70% porosity we expect, therefore, to find either or both of the phases continuous or partially continuous.

The model assumed for sintered polycrystals of barium titanate is, therefore, one of spherical air pores dispersed in a continuous medium of barium titanate for porosities up to about 35%, for which region $u=2\epsilon_1$, and for porosities from 35% to about 45% it is assumed that the barium titanate is wholly continuous and the air partially continuous, the *u* factor being indeterminate. Beyond about 45% porosity the barium titanate is discontinuous and the air continuous since the particles cannot be sintered together at such high porosity by this method. The *u*-factor in this region, assuming spherical particles not too close together, is $2\epsilon_2$ as stated in table 2. The "high porosity" formula (6) should, however, be applicable to systems such as $BaTiO_3$/wax, in which the low-permittivity wax forms the continuous phase.

It is not possible to compute a generalized *u*-factor for the region 30% to 70% porosity and hence to deduce a mixture law, but if normal methods of preparation are used one would expect a smooth transition between the Wiener formulae (4) and (6), which in turn implies smooth transition between the corresponding *u*-factors, $2\epsilon_1$ and $2\epsilon_2$.

Little can be said about the precise nature of the variation in *u* over the intermediate region, as it will depend on the exact conditions of preparation of the material, e.g. the density distribution within the compacted material before sintering, and more especially on the physical forces acting on the particles during the sintering itself process. In this region there would appear to be little hope of deducing a rigorous mixture law, and the only formula having any practical utility would be one giving an empirical relationship between the *u*-factor and composition.

ACKNOWLEDGMENTS

The authors wish to thank Mr. J. A. M. van Moll and the directors of Philips Electrical Ltd. (formerly Philips Lamps Ltd.) for permission to publish this paper.

REFERENCES

LICHTENECKER, 1926. *Phys. Z.*, **27**, 115.
POLDER and VAN SANTEN, 1946. *Physica*, **12**, 257.
RUSHMAN and STRIVENS, 1946. *Trans. Faraday Soc.*, **42** A. 231.
WIENER, 1904. *Phys. Z.*, **5**, 332.
WIENER, 1909. *Leipzig Ber.*, **61**, 113; 1910. *Ibid.*, **62**, 256.

DIFFUSION PUMPS: A CRITICAL DISCUSSION OF EXISTING THEORIES

BY D. G. AVERY* AND R. WITTY,
Metropolitan-Vickers Electrical Co. Ltd., Manchester, 17

MS. received 8 *May* 1947; *read at Science Meeting on* 31 *October* 1947

ABSTRACT. The original work of Gaede and others on diffusion pumps is briefly described. A critical discussion of this work leads to the formulation of a more complete theory of the action of the diffusion pump, and this is used to explain the practical characteristics of a simple form of modern diffusion pump.

§ 1. INTRODUCTION

THE main contributions to the theory of the diffusion pump between 1915 and 1923 were those made by Gaede (1915, 1923), Langmuir (1916) and Crawford (1917). Since that time several critical surveys of these theories (see General Bibliography) have been made, but it is felt that none of them have reached definite conclusions. The purpose of this paper is to correlate the previous work mentioned above and then put forward a theory of the diffusion pump which will satisfactorily explain the known phenomena of the pumps. In this paper considerations are limited to cases where the working pressure of the pump is less than 10^{-3} mm. of mercury.

§ 2. ORIGINAL THEORIES

(a) *Gaede's theory of the diffusion pump*

The original work which led to the development of the Gaede air-diffusion pump (Gaede, 1915) was concerned with diffusion processes due to gas and vapour counter-flowing through a tube under the conditions shown in figure 1.

---

* Now at Royal Holloway College.

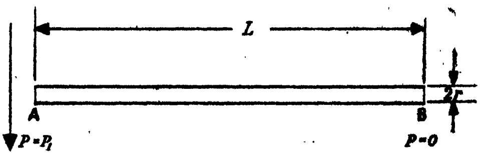


Figure 1.

Suppose that at the two ends A and B of the tube, of radius r and length L, the vapour pressure P has the values P_1 and zero, and the gas pressure p has values zero and p_2, respectively. When these conditions are satisfied (for example by creating a gas-free vapour stream in the direction of the arrow at A, and placing a condenser at B), then the volume of gas V flowing per second from B to A is given by

$$V=\frac{1}{L}\cdot\frac{\pi r^3}{2v_1}\cdot\exp\frac{-rP_1}{1520Dv_2}, \qquad \ldots\ldots(1)$$

where v_1 and v_2 are the coefficients of external friction of gas and vapour respectively, and D is the coefficient of diffusion for the gas and vapour concerned. It should be noted that V is greatest when rP_1 is least and L is small. Thus if the vapour pressure P_1 is high, the radius r must be very small, and the volume V may be increased by using a number of similar tubes in parallel; alternatively, if the vapour pressure P_1 is very low, then r may be large. Since at low gas pressures for gas and vapour counter-flowing in a tube, the vapour pressure P_1 completely determines the value of the mean free path, λ, of the gas molecules in the vapour, it follows that the volume V depends on λ.

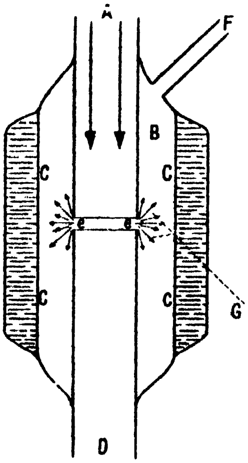


Figure 2.

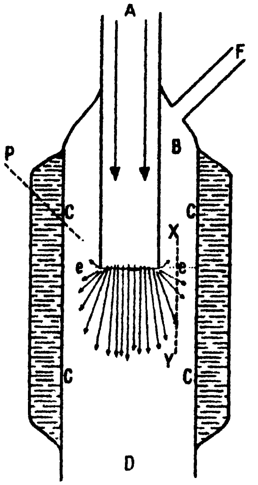


Figure 3.

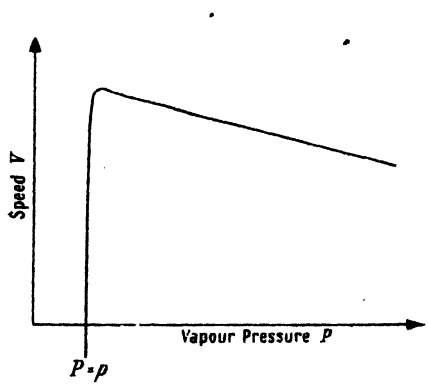


Figure 4.

Gaede (1915, 1923) used these ideas as a basis for the construction of the pumps of the type shown in figure 2. In these pumps the diffusion processes take place through the gap e. A stream of gas-free vapour passes in the direction shown from A to D past this gap and is condensed at D in a space already evacuated by a backing pump. Some of this vapour passes out through e into the space B, emerging in a brush-like formation to be condensed at the cooled walls C. Air from the vessel to be exhausted enters the system at F and, after diffusing through this cloud of vapour, enters the stream at e. Thus the conditions prescribed in the previous paragraph are satisfied.

It has been shown that the rate of diffusion depends on the product rP_1, where in this case r is the gap-width. Therefore, since in the region G the mean free path λ of the gas molecules in the vapour issuing from the gap is inversely proportional to P_1, the value of the factor r/λ controls the rate at which gas passes through the pump. If λ is small, then the probability that a gas molecule will penetrate through the vapour from G and enter the stream at e is small, and, in the limiting case, the issuing vapour sweeps the gas back from the region of the gap. On the other hand, if λ is large and r/λ small, then the gas easily enters the stream, suffering very few adverse collisions with vapour molecules before reaching e. If P_1 is made smaller, and λ in consequence larger, then it follows that the gap may be made larger and an increase in speed obtained. An upper limit is placed on the free path length since, if this is very great, and the vapour pressure correspondingly very low, the gas that has passed through the gap is no longer carried to the fore-vacuum at D. Gaede stated as a condition for optimum working that the mean free path λ should be of the same order of magnitude as the gap-width r.

The pressure exerted by the vapour at G on the air in the vessel B, termed the *diffusion back pressure*, is given by

$$\Delta p = p_2\left(1 - \exp\frac{-rP_1}{1520Dv_2}\right). \qquad \ldots\ldots(2)$$

The exponential term in this equation must be large, i.e. rP_1 must be small, as shown above, for Δp to be small. If rP_1 is large, then Δp is large and no gas diffuses into the vapour stream.

Gaede calculated, on the basis of the kinetic theory, that, for this form of pump, the volume of gas passing per second from F to D is given by

$$V = \alpha q/2\pi\rho_1. \qquad \ldots\ldots(3)$$

In this equation q is the area of the gap, ρ_1 the density of the gas at the working temperature under unit pressure, and α a function of r/λ which decreases as r/λ increases, or as λ decreases. V is independent of p_2, the pressure at F, so that there is no theoretical limit to the vacuum which may be obtained.

Gaede further showed (1923) that equation (3) should be modified to

$$V = k\frac{\alpha q}{2\pi\rho_1}, \qquad \ldots\ldots(4)$$

where k is a constant taking into account the fact that the vapour stream may not be able to remove all the gas arriving at e. He considered that the layers of vapour nearest the gap e become saturated with gas and can no longer entrain all the gas arriving. He claimed that the amount of gas which any stream can "absorb" is greatest when the individual points of the vapour stream remain exposed to the gas for the shortest possible time, so that, for a given vapour stream, k is large when r is small, or, for a given r, k can be increased by increasing the velocity of the stream, e.g. by using driving nozzles. In the case of pumps of the type shown in figure 3, with the diffusion gap at e, where the density of the stream increases rapidly with the depth of penetration along a line such as XY, Gaede suggested that the surface layers may become more easily saturated and the value of k correspondingly less.

Equation (4) gives the volume of gas passing per second from F to D (figure 2); it is not, however, a true equation for the pumping speed. Gaede quoted the complete equation as

$$V=\frac{q}{2\pi p_1}\left(k\alpha-\beta\frac{p_1}{p_2}\right), \qquad \ldots\ldots(5)$$

where p_1 is the pressure existing in D, the second term accounting for gas back-diffusing from D to F. The quantity β decreases rapidly with increasing vapour pressure and stream speed, and so, for any particular value of p_1, the speed is independent of this value once P_1 has exceeded a value such that $\beta=0$. On the other hand, if P_1 is equal to or less than p_1, then the stream can no longer hold back the pressure in D and the pump ceases to function.

If the speed is plotted aainst the vapour pressure, the curve shown in figure 4 is obtained. The pump has zero, or negative, speed so long as $P_1<p_1$, and immediately rises to a maximum as soon as P_1 equals p_1. Thereafter the speed slowly declines with increasing P_1, since λ and α both decrease.

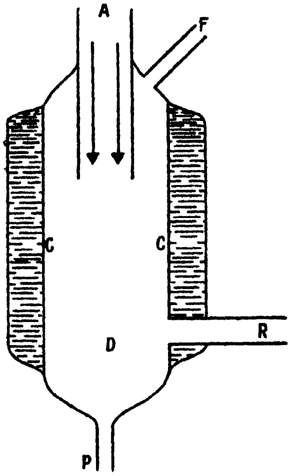


Figure 5.

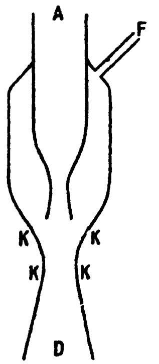


Figure 6

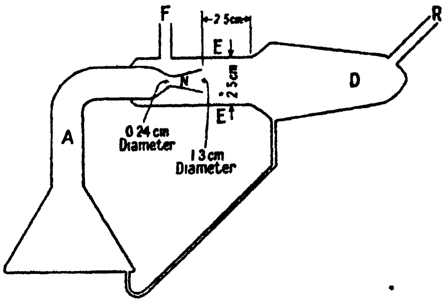


Figure 7.

Gaede concluded the theoretical discussion by pointing out that if the speed of the stream is increased, the speed of pumping is increased, not only due to the increase in k, already mentioned, but also, especially in the type of pump shown in figure 3, due to the increase in α as a result of the reduction of the chances of a vapour molecule "back-streaming" through the gap e into the space B.

Several interesting experimental confirmations of Gaede's theory have been made by Molthan (1925, 1926).

(*b*) *Langmuir's theory of the condensation pump*

A typical Langmuir pump (Langmuir, 1916) is shown in figure 5. Mercury vapour passes down the tube A into the larger tube D, the walls of which are cooled by the condenser C. The vapour, on leaving the tube A, spreads out and is condensed on the walls, whence it returns to the boiler by way of the tube P. The high-vacuum side of the pump is at F, the backing pump being connected at R, Langmuir considered the action of all high-vacuum vapour pumps to be divided into two separate processes, viz.,

(1) the process by which the gas enters the blast of vapour; and

(2) the process by which the blast carries the gas along into a condensing chamber from which it cannot return to the vessel being exhausted.

The second of these processes is practically the same for Gaede's diffusion pumps, steam-ejector pumps, and the pump under consideration. Langmuir considered that the effectiveness of the Gaede pump lay in the success of the second process, and that the limitations were those imposed by the first, that is by the restrictions on the size of the diffusion orifice.

Langmuir proceeded to consider ejector pumps; in such pumps the gas is drawn into the vapour because the pressure in the jet at the point where the gas enters it is lower than the pressure of the surrounding atmosphere, that is, these pumps utilize the Bernouilli effect. This effect cannot be directly utilized in the production of high vacua, since the pressure in the jet must always be considerable and the jet will expand laterally, as predicted by the kinetic theory.

If an ejector pump such as that shown in figure 6 is assumed to be working with a high vacuum on the intake side, then the vapour blast from A will expand laterally to strike the walls at K. The wall will rapidly assume the temperature of the vapour and vapour molecules will leave the walls at K with high velocities in all directions and thus completely destroy the pumping effect. This Langmuir confirmed by experiment.

Langmuir then went on to consider that if these molecules could be prevented from leaving K after first hitting the walls, then their downward-velocity component could be utilized to transfer gas from the high-vacuum side to the backing pump. If the process by which the vapour molecules passed from K into the high-vacuum side was one of reflection, then there was no way out of the difficulty, but Langmuir's previous work had shown that the effect was almost entirely one of condensation followed by random re-emission. Accordingly, he concluded that if the walls of the pump were kept cool, as in figure 5, by the cooler C, then a pumping action could be obtained.

This was borne out by experiment, and Langmuir (1916) summarized the action of his pump as follows:—"In these pumps a blast of mercury vapour carries the gas into a condenser. This action is similar to that in a steam ejector and in a Gaede diffusion pump. The method by which the gas is brought into the mercury-vapour blast in the condensation pump is based on a new principle which is essentially different from that employed in the steam ejector or Gaede diffusion pump. In the new pumps the gas to be exhausted is caught by the blast of vapour and is forced by gas friction to travel along a cooled surface. By maintaining this surface at such a low temperature that the condensed mercury does not re-evaporate at an appreciable rate, it is possible to keep the mercury vapour from escaping into the vessel being exhausted. The action of this pump therefore depends primarily upon the fact that all the atoms of mercury striking a mercury-covered surface are condensed (no matter what the temperature), instead of even a fraction of them being reflected from the surface. It is for this reason that the term 'condensation pump' is proposed."

From these considerations it can be seen that such a pump should have no limiting vacuum, since a limit could only be imposed by the back-diffusion of gas through the stream, and calculations show that the chance of a gas molecule finding its way back through the stream in a typical pump is only 1 in 10^{20}.

(c) *Crawford's theory of the parallel jet pump*

Crawford (1917) considered that the primary obstacle to the successful operation of vapour pumps was the fact that the vapour stream dispersed when surrounded by a high vacuum, and did not satisfactorily entrain the gas from the vessel to be exhausted.

He considered that if a jet could be produced, the molecules of which had equal and parallel velocities, one with another, then such a jet of vapour should be collision-free and should not materially disperse in a high vacuum. Also it would be a very efficient pumping agent since gas could readily enter the stream in a direction perpendicular to its axis, but gas molecules once in the stream would be carried along with it in the direction of its motion. He further considered that, even if collisions did occur in such a jet, the stream would not necessarily be dispersed since only the direction of relative velocities, and not the velocity of the centre of gravity of the whole mass, would be altered by such collisions.

These considerations led Crawford to design pumps of the type shown in figure 7. The form of nozzle N was settled by the considerations of steam-engineering practice, the point of minimum pressure and maximum velocity existing in the diverging portion of the nozzle. With such a pump, Crawford stated that no diffusion slit or condensing surface was necessary. The tube E could be artificially heated without seriously affecting the performance, this suggesting that the jet itself re-entrained and expelled vapour molecules diffusely emitted from the walls. He calculated that in a typical jet a vapour molecule would suffer, on an average, six collisions between leaving the nozzle and being condensed in D.

If the jet-density exceeded a well-defined limit, he found that the pump ceased to work, and he explained this by saying that "this limit is established by the density of the dispersing fringe, which is probably proportional to the density of the jet and occurs at the point where the mean free path of the gas molecules entering the fringe becomes less than the total depth of the fringe".

§ 3. DESIGN DETAILS

(a) *Gaede*

Gaede's first designs of pumps (1915) were all based directly on the fundamental arrangement shown in figure 2. A thermometer was usually inserted to measure the temperature of the vapour, and hence the vapour pressure, at some point near the gap, since the heating adjustment, for the correct value of P, was very critical (cf. figure 4). With these pumps he obtained an optimum speed of 80 cc./sec. with air.

Following on from the theoretical considerations discussed, he added a driving nozzle to increase the velocity of the vapour stream (1923), giving the arrangement shown in figure 8. With this pump he obtained a speed of 0·25 litres/sec., using the same gap-width as in the first pump mentioned above. Since the effective mean free path of the vapour molecules in the stream had been increased by the use of the nozzle, it was possible to increase the gap-width and, using an annular nozzle ring shaped in section, to give a high-velocity stream, he obtained a speed of 1·5 litres/sec. using the same tube diameters as those in figure 8. A larger pump, in which the tube B had an internal diameter of 5 cm., gave speeds of up to 6 litres/sec. He found that when a nozzle was inserted, the adjustment of vapour temperature had no longer such a critical effect on the speed.

Gaede's latest type (1923) was a multi-stage pump developed from pumps of the type shown in figure 3, again utilizing a driving nozzle. The first stage of the pump had a nozzle form as shown in figure 9 (*a*), the diffusion orifice again being at e. This stage was backed by two or more stages of the general type shown in figure 9 (*b*). With such a pump Gaede obtained speeds of 60 litres/sec., with a value of $k\alpha$ in equations (4) and (5) of 0·4.

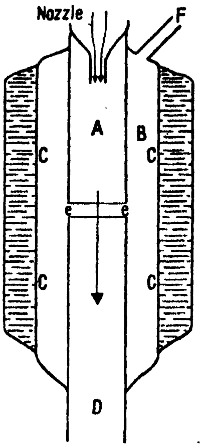
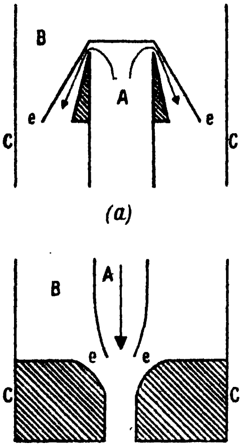


Figure 8.



(*a*)

(*b*)

Figure 9.

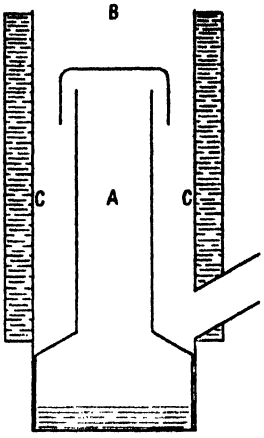


Figure 10.

In all cases, when working with air, he found that the speed and ultimate pressure were independent of the backing pressure below a certain critical value of that pressure.

(b) *Langmuir*

The first design of condensation pump has already been shown in figure 5; it should be noted that this system is identical with the system shown in figure 3. In these pumps, as in the later Gaede pumps, the adjustment of temperature was by no means critical. Langmuir observed critical-backing pressure effects, and further found that the value of the critical-backing pressure increased rapidly with increasing heat input.

Langmuir's major contribution from the design point of view was the construction of all-metal vertical pumps of the type shown in figure 10. Using one of these pumps, with a casing internal diameter of 7 cm., he obtained a speed of 3 litres/sec.

(c) *Crawford*

The design of Crawford's pump (1917) was discussed in § 2 (c), the nozzle being designed from steam-engineering principles. A dimensional drawing of one of his pumps is shown in figure 7. The curves of speed against boiler pressure show a definite, though not critical, optimum boiler pressure, giving, for a typical pump, a speed of 640 cc./sec. Other pumps had speeds as high as 1·2 litres/sec.*

§ 4. CRITICAL DISCUSSION

From the theories developed in the preceding two sections it appears possible to give a fairly complete description of the action of these types of high-vacuum vapour pumps. It can be seen that a successful pump must be constructed so as to fulfil satisfactorily two prime conditions; first, the gas on the fine side must have easy access to the pumping stream, so that it may be removed quickly to the backing side; secondly, the pump must act as an efficient pressure seal so as to give as low a pressure as possible on the fine side whilst operating against as high a backing pressure as possible. Following Langmuir, then, the action of a vapour pump may be divided into two parts:

(1) the entry of the gas into the stream;
(2) the removal of the gas to the backing side.

Back diffusion, critical backing pressure

The second of these processes is practically the same in all pumps. Once the incoming gas has been entrained in the stream, it is forced to move in the direction of the stream towards the backing side by large numbers of collisions with the comparatively dense, fast-moving, heavy vapour-molecules. These same collision processes prevent almost all the gas in the backing side from forcing its way back into regions of lower pressure. In all pumps the operating characteristics show that for heavy gases, and at all values of the backing pressure below a certain critical value, only a very small proportion of the gas on the backing side "back-diffuses" into the fine side space. Above this critical backing pressure the gas pressure on the backing side is greater than the pressure exerted by the stream, and the stream is forced back, with a consequent destruction of the pressure-sealing effect.

In the case of light gases greater amounts of gas back-diffuse at backing pressures below the critical, and since for any gas the amount back-diffusing (at pressures below the critical) is clearly dependent on that pressure, for light gases the critical effect is less marked.

Fundamental condition for efficient pumping

Process (1) is concerned with the entry of the gas into the pumping stream, that is with the passage of the gas from a point such as P, figure 3, into the pumping stream through the gap e. Now in any pump a certain number of vapour molecules will escape from the stream and proceed in the direction from e to P, and any gas molecule entering the stream has to pass through a "cloud" of these "backstreaming" vapour molecules, which are moving in straight-line paths from points in the pumping stream. Obviously the gas will enter the stream most readily *when the mean free path of the gas molecules in the cloud of backstreaming vapour is as large as*

---

* In comparing the performances of these various types of pump it would be more instructive to give the values of $k\alpha$, since the areas q of the several pumps are not the same. Unfortunately, however, there are not sufficient published data to enable this to be done.

possible. Under these conditions, the number of collisions suffered by a gas molecule with the vapour is a minimum, which is desirable since each collision will result in the gas molecule largely assuming the direction of the velocity of the backstreaming molecule. *Therefore, for process* (1) *to be carried out efficiently in any pump, the backstreaming must be reduced to a minimum.*

Causes and elimination of backstreaming

Backstreaming is caused by collisions between the vapour molecules in the pumping stream occurring whilst the stream is exposed to the fine-side space, resulting in vapour molecules leaving the stream in directions contrary to the pumping direction. The density of the backstreaming molecules is proportional to the number of collisions so occurring, and, therefore, for a given speed of stream and size of diffusion orifice, it is proportional to the pressure and, therefore, the temperature of the molecules in the stream. *The only important differences between the various types of pump considered are those between the methods employed to reduce the backstreaming of the vapour molecules to a minimum.*

In Gaede's pump, figure 2, backstreaming was reduced by adjusting the temperature (and thus the density) of the stream and the width of the slit, so that the number of collisions occurring between the molecules of the stream whilst it was exposed to the fine side was a minimum. A lower limit was placed on the stream density by the considerations of process (2). When the speed of the vapour stream was artificially increased, the time during which any portion of the stream was exposed was reduced, and hence it was possible to increase the slit-width without fear of increased backstreaming.

In Langmuir's pump of the type shown in figure 5, backstreaming was reduced by constructing the system so that a vapour molecule had to turn through 180° to the direction of motion of the stream before it was proceeding directly contrary to the pumping direction, as opposed to 90° in the Gaede pump, and, since the walls of these pumps were vigorously cooled, there was very little random re-emission of vapour molecules from the walls after they had once passed across the gap. Some control on the number of collisions occurring in the stream whilst it was exposed to the fine side, had, of course, still to be maintained by the correct adjustment of the vapour temperature, but since any collision was less likely to have a serious effect on the action of the pump, this adjustment was not nearly so critical as in the Gaede pump.

Crawford constructed his pump so that the stream was directed in the pumping direction and so that the vapour molecules should have high velocities with near parallel directions. That the use of a high-velocity stream reduces the chance of a molecule backstreaming follows from the fact that for a molecule to stream back it must have a thermal velocity greater than the streaming velocity. Thus, if the streaming velocity is greater, fewer molecules will have thermal velocities greater than the streaming velocity. Gaede realized this, and, when he increased the speed of the stream by the use of driving nozzles, he was able to use a wider slit and obtain greater speeds, since the vapour molecules streamed back less readily.

Drift vs condensation

Gaede stated that all vapour pumps depended for their successful action on the fact that the gas enters the system through a diffusion diaphragm, defining the latter as an orifice through which gas and vapour were flowing in opposite senses and in which the mean free path of the gas molecules in the vapour was of the same order as the linear dimensions of the orifice. We have seen that this is true in a sense, since in all pumps the gas must enter the stream through a cloud of backstreaming vapour, and the mean free path of the gas molecules in this vapour must be large.

In this way all vapour pumps may be said to operate on the Gaede principle since they will only operate satisfactorily when backstreaming has been reduced until the mean free path of the incoming gas molecules in it is large. In the Gaede pump, backstreaming is reduced by adjusting conditions to give the correct stream density–slit-width relation, and in the Langmuir by directing the stream.

The effect of excessive heating

The critical effect of backstreaming on the entry of gas into the stream is confirmed by observations on the effect of increasing the temperature of the vapour stream without increasing its speed. In all pumps, as the vapour molecules get hotter and collisions more frequent, the amount of backstreaming increases and the speed of the pump falls off. Gaede has accounted for this in his formula for the speed of the pump, the coefficient α decreasing with increasing vapour pressure and temperature. In the Langmuir pump the speed falls off with increasing heat input (after a maximum has been passed), and if the walls are not sufficiently cooled to prevent emission from them, this will also cause a fall in the speed of pumping. Crawford finds that the speed of his pump falls off sharply with increasing heat input after a maximum value, but says that this is due to the density of the dispersing fringe exceeding a certain value. As Gaede points out, however, it is more likely that the fall in speed is due to an increased number of collisions in the stream resulting in increased backstreaming.

Absorptivity

In addition to the effect of backstreaming on the entry of gas into the stream, there is the secondary factor which Gaede calls " absorptivity ". This factor takes into account the fact that the vapour stream is not capable of entraining and removing all the gas molecules which reach it. Gaede suggests that this is due to the saturating of the surface layers of the stream with gas, and he gives experimental evidence in confirmation of this, and suggests that in pumps of the Langmuir type, where the stream density increases rapidly with depth of penetration, saturation may occur more readily than in his own pumps, which utilize a stream of uniform density. Beyond this little is known of the effect, and none of the other authors quoted refer to it, though it seems highly probable that some such effect exists.

§ 5. DEVELOPMENTS SINCE 1923

The theoretical conclusions given above were developed from work done by Gaede, Langmuir and Crawford between 1915 and 1923 (*loc. cit.*), and there has been no further published work of a sufficiently fundamental nature since that date to affect the conclusions. Most of the more recent developments in the field have been in design, with the object of reducing the quantity of backstreaming vapour to a minimum. Of the papers published between 1923 and 1939, those of Ho (1932 a and b) require special mention in that he introduced the " Ho coefficient " for the efficiency of a vacuum pump. This coefficient is defined as the ratio of pump speed to the theoretical (Knudsen) admittance of the annular gap round the nozzle of a pump. Nearly all the pumps developed used the directed stream of the Langmuir pump and utilized nozzle forms to obtain high-velocity pumping streams as Crawford did. The first of these was that designed byGaede (1923), already referred to (figure 9 (*a*) and (*b*)), and which was subsequently improved (Gaede and Keesom, 1929).

The most recent developments are due to workers in America and are well summarized by Hickman in a recent paper (1940). The high-speed jets of Embree (U.S. Pat. 2,150,676) and Stallman (Brit. Pat. 565,455) are both designed to give a high-velocity downward-directed stream to reduce backstreaming and increase absorptivity, resulting in higher speeds for the reasons given above.

In a recent paper, Alexander (1946) described a successful high-vacuum pump, the design of which was based on the desirability of reducing the backstreaming to a minimum. In the discussion of the theory of the pump, however, several misconceptions arise which are commonly met with in such discussions. Alexander stated Gaede's condition for the operation of a diffusion pump as being that the mean free path of the vapour molecules in the stream shall be of the order of the width of the orifice. This statement, which is to be found in other books and papers on the subject, is incorrect.* The correct statement is that *the mean free path of the gas molecules in the backstreaming vapour* shall not be less than the dimensions of the orifice. In a Gaede type of diffusion pump (figure 2), for the fundamental Gaede condition to hold, that quoted by Alexander must hold as well, but, as has been shown, in diffusion pumps of the Langmuir or Crawford types, whilst in these pumps, too, the mean free path of the incoming gas in the backstreaming vapour must be of the order of the dimensions of the orifice, the pressure of the vapour in the stream can be considerably greater than that which would be satisfactory in a Gaede pump. From this it can be seen that the correct inference to be drawn from Alexander's calculations of the amount of backstreaming, which show that the density of the backstreaming vapour is very low, is that his pump does satisfy Gaede's principle, and is a " diffusion " pump. The fact that the ratio of the speed of the pump to the area of the orifice is not dependent on the throat width, as might be expected from a first consideration of Gaede's theory, is easily explained by saying that the jet of his pump reduces backstreaming so effectively that even at the largest throat-widths met with it is still insufficient materially to affect the speed of the pump.

A paper shortly to be published by Blears and Hill † deals with the performance of the diffusion pump with light gases.

§ 6. SUMMARY

For any diffusion pump to be successful it must satisfy two conditions :—

(1) The incoming gas must be able to enter the pumping stream easily.

(2) The pumping stream must effectively prevent the gas in the backing side from passing into the high vacuum side.

---

* This misconception probably arose because in Gaede's first paper (1915) he himself did not make it clear that the important factor was the mean free path of the gas molecules in the backstreaming vapour molecules. This was clarified in his 1923 paper, but for some reason the significance of this paper has been largely overlooked. In fact, the present authors were not familiar with it until recently and are very grateful to Mr. H. Griffiths of London for drawing their attention to the importance of Gaede's 1923 paper.

† Research Department, Metropolitan-Vickers Electrical Co. Ltd.

The first condition was stated by Gaede as: "the mean free path of the gas molecules entering the pump in the vapour molecules backstreaming from the pumping stream must not be less than the dimensions of the orifice". This condition may be simply stated as "backstreaming must be reduced to a minimum".

In the three major types of pump this is done in three different ways:

(i) In the Gaede-type pump (figure 2) backstreaming is reduced by increasing the mean free path of the vapour molecules in the stream. If this mean free path is increased until its value is greater than the linear dimensions of the orifice, then condition (2) breaks down. The pump operates most satisfactorily when the mean free path of the vapour molecules in the stream is of the order of the linear dimensions of the orifice.

(ii) In the Langmuir-type pump (figure 5), backstreaming is reduced by directing the stream in the pumping direction and condensing it on impact with the walls. It must be remembered, as a second-order effect, that the mean free path of the vapour molecules in the stream must not be too small a fraction of the dimension of the orifice.

(iii) In the Crawford-type pump (figure 7), backstreaming is reduced by directing the stream and giving the vapour molecules a very high velocity in the pumping direction.

In any pump, increasing the velocity of the vapour stream tends to increase the effective mean free path of the vapour molecules in the stream, and so further reduce backstreaming.

Most modern diffusion pumps are a combination of the Langmuir and Crawford types.

§ 7. EXPERIMENTAL CHARACTERISTICS

It can now be shown how the working characteristics of a simple form of modern diffusion pump may be explained qualitatively by means of the theory developed above. The characteristics considered are those of a small diffusion pump, a "Metrovac" type 03 (figure 11) having a single umbrella-type jet and operating with "Apiezon B" oil. This form of jet is typical of those in common use today.

It must be remembered that, whilst the general form of the characteristics given is the same for all pumps and may be correlated with the theory, any quantitative values quoted are critically dependent on the particular size of pump and type of working fluid, and cannot be used as checks for the theory without a more precise knowledge of these factors.

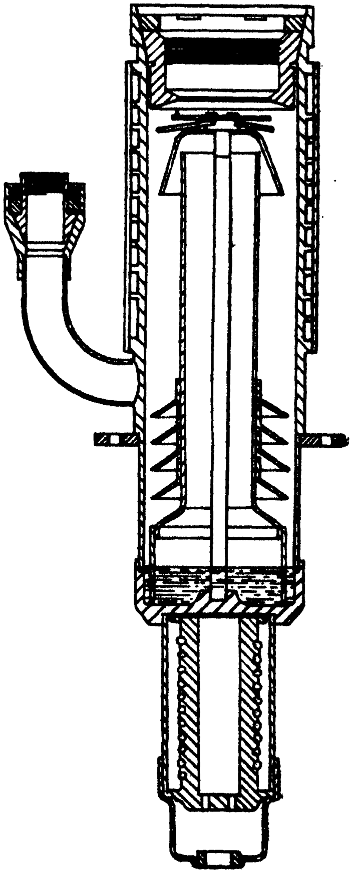


Figure 11.

1. *The relation between speed and heater wattage*

The curve obtained by plotting values of the speed, at a very low backing pressure ($<1\mu$), against the heater wattage is shown in figure 12. The general form of this characteristic shows an optimum value of heater wattage, the maximum in the speed values not being sharply defined. The explanation follows directly from the theory. At low heater wattages, where the vapour pressure is just less than or equal to the backing pressure employed, the stream is insufficient to seal the pump, and thus it has zero speed. As the heater wattage is increased the sealing becomes more efficient. This improvement is offset by an increase in backstreaming; this follows since the increased temperature of the stream causes an increased number of collisions between the vapour molecules of the stream. For a given backing pressure, once the pump acts as an efficient seal, an increase in the heater wattage merely serves to increase the backstreaming and hence decrease the speed, the curve thus passing through a maximum. Obviously, for a greater backing pressure (dotted curve in figure 12), more heater watts are required to seal the pump. Consequently the optimum value of the speed is less than with the lower backing pressure, since it occurs at a higher vapour temperature.

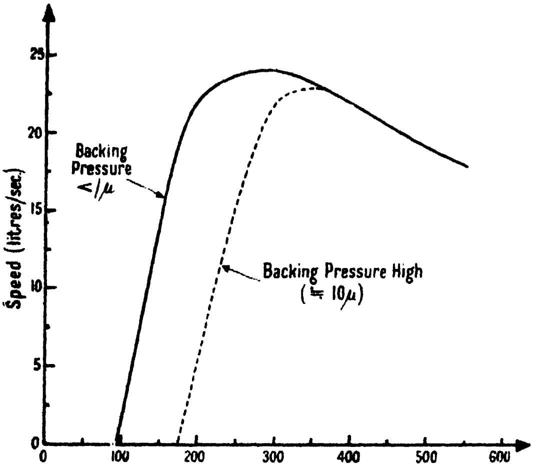


Figure 12.

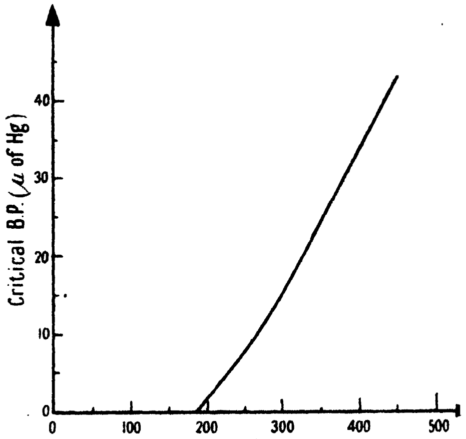


Figure 13.

2. *Critical backing pressure—heater-watts relation*

The relation between critical backing pressure and heater watts is shown in figure 13. At low-heater wattages, where the vapour pressure is very small, the critical backing pressure is practically zero. As the heater watts are increased, so the critical backing pressure increases, the relation becoming nearly linear at higher wattages. As the wattage increases, the temperature of the stream increases, and hence, also, its potential energy. Again, for a given throat-width in the pump nozzle, the velocity of the stream increases, due to a greater pressure difference across the nozzle, and hence the kinetic energy of the stream also increases. The stream, with total energy increased, is thus able to seal off a larger pressure on the backing side.

3. *The relation between fine side pressure and backing pressure and between speed and backing pressure*

As has been discussed, the amount of gas back-diffusing through the stream, for heavy gases, is practically zero for backing pressures below the critical, and above the critical this amount increases very rapidly. This satisfactorily explains the shape of the experimental curve shown in figure 14.

Since the speed of the pump may be regarded as the net rate of transfer of gas in the pumping direction—that is, the total amount passed in the pumping direction minus that diffusing back—the shape of the curve of speed against backing pressure may easily be derived from the curve above. An experimental curve is shown in figure 14, the speed remaining independent of the backing pressure until the critical value is passed, when the speed falls off sharply.

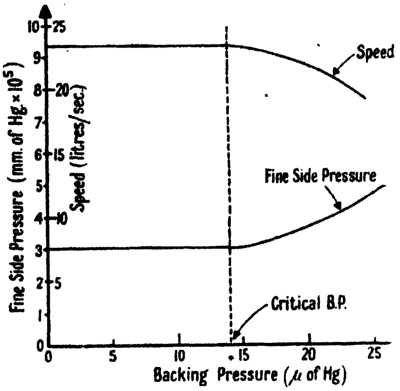


Figure 14.

4. *The relation between speed and fine side pressure*

We have seen that, in the absence of back-diffusion of gas, the pumps discussed have no limiting vacuum, and hence, under these conditions, the speed is independent of the fine-side gas pressure. If the back diffusion is not zero, then the speed (regarded as the net transfer of gas in the pumping direction) falls to zero when the transfer of gas in the pumping direction is equal to that in the reverse direction. In this condition the pump merely passes gas from the fine to the backing side, whence it returns by diffusing back to the fine side. When the amount of gas on the fine side due to back diffusion is negligible compared with that due to other sources, then the speed of the pump is independent of the fine-side pressure.

§ 8. CONCLUSION

It has been shown that the modern diffusion pump is an outcome of the original pioneer work of Gaede, Langmuir and Crawford. When observed in the true perspective, it is seen that the contributions made by these three were not competitive, but rather were complementary, the best pumps depending on advances made by all three.

We have not tried to apportion the credit for any particular part of the theory, but rather to synthesize a complete theory based on all the data available.

ACKNOWLEDGMENTS

The authors wish to express their thanks to Mr. T. M. Hodson for the translation of Gaede's papers and also to Sir Arthur P. M. Fleming, C.B.E., D.Eng., Director of Research and Education, and Mr. B. G. Churcher, M.Sc., M.I.E.E., Manager of Research Dept., Metropolitan-Vickers Electrical Co. Ltd., for permission to publish this paper.

BIBLIOGRAPHY

ALEXANDER, 1946. *J. Sci. Instrum.*, **23**, 11.
CRAWFORD, 1917. *Phys. Rev.*, **10**, 557.
GAEDE, 1915. *Ann. der Phys.*, **46**, 357.
GAEDE, 1923. *Z. tech. Phys.*, **4**, 348.
GAEDE and KEESOM, 1929. *Zeit. f. Instrum.*, **49**, 298.
HICKMAN, 1940. *J. Appl. Phys.*, **11**, 303.
HO, 1932 a. *Rev. Sci. Instrum.*, **3**, 133.
HO, 1932 b. *Amer. Phys. Soc. Proc.*, Jan.-June 1932, 386.
LANGMUIR, 1916. *Gen. Elect. Rev.*, **19**, 1060.
MOLTHAN, 1925. *Phys. Z.*, **26**, 712.
MOLTHAN, 1926. *Z. Phys.*, **39**, 1.

GENERAL BIBLIOGRAPHY

DUNOYER, 1924. *Vacuum Practice.*
GOETZ, A., 1926. *Physik und Technik des Hochvakuums* (Chaps. 1 and 2).
NEWMAN, 1925. *Production and Measurement of Low Pressure* (Chap. 5).
WITTY, R., 1945. *J. Sci. Instrum.*, **22**, 201.
The Dictionary of Applied Physics.

THE EFFECT OF TEMPERATURE ON THE THERMAL CONDUCTIVITY AND VISCOSITY OF LIQUIDS

BY R. EISENSCHITZ,
University College, London

Communicated by E. N. da C. Andrade, F.R.S. ; *MS. received* 26 *August* 1947

ABSTRACT. The theory of Brownian movement is applied to the 'cell model' of liquids. The Smoluchowski equation for the density distribution is solved under conditions corresponding to laminar flow and conditions corresponding to inhomogeneous temperature. The flow of momentum and the flow of energy are calculated. The result indicates that the coefficient of viscosity varies exponentially with the inverse temperature and that the coefficient of thermal conductivity is proportional to the absolute temperature.

§ 1. INTRODUCTION

As early as 1885 Graetz noticed that the effect of temperature on the viscosity and thermal conductivity of liquids is of opposite sign and different magnitude, whereas in gases these two transport coefficients are known to be almost proportional to each other. Graetz argued that the intermolecular forces have only an indirect influence on the mechanism of heat transfer, but exercise resistance against deformation.

The relation between temperature and viscosity has subsequently been found to follow an exponential law, the significance of which for the kinetic theory of liquids was first recognized by Andrade (1934). Little attention has, however, been given to the differences between the two transport phenomena, which are so marked that even a very rough model of the liquid state should indicate their origin.

The model of a liquid which is used in this paper is chosen from the point of view of mathematical simplicity rather than the close expression of experimental data. A model of this kind is the 'cell model', representing a molecule in the field of force of its neighbours. The motion of the representative molecule may be regarded as Brownian movement and, in particular, the distribution in space be supposed to be governed by the Smoluchowski equation.

Although this is a crude model, it is in accordance with Kirkwood's theory (1946) of irreversible processes, in which he succeeds in deriving irreversible changes of the average state of motion of one molecule or of a pair of representative molecules from the reversible mechanics of the totality of molecules. He shows that the distribution of a pair of molecules in coordinates and momenta is determined by an equation of the type that applies to the Brownian movement of colloid particles, and he derives the friction constant appearing in this equation from molecular data. His theory of viscosity and thermal conductivity requires the determination of the distribution in phase of a pair of molecules which depends on the solution of a partial differential equation with twelve independent variables.

The comprehensive and systematic kinetic theory of liquids being worked out by Born and Green (1946, 47, Green, 1947) has already given valuable results, but by the nature of the rigorous methods employed is not likely to lead to simple formulae.

By considering the distribution of one representative molecule instead of a pair—that is by means of the cell method—fairly good results are obtained in the field of equilibrium theory. The differential equation in the theory of Brownian movement determining the distribution in momenta and coordinates may be replaced by the Smoluchowski equation in coordinates only, if the friction constant is sufficiently high.

A method similar to that used in this paper has been employed by Kramers (1940) for calculating the rate of transitions of particles over a potential barrier.

The cell model will accordingly be used for calculating the transport coefficients of the liquid in terms of the friction constant and of the particular law of force within the cell. The latter is assumed to be proportional to the distance from the centre, i.e. to be derived from a parabolic potential field increasing from the centre to the surface of the cell and remaining at constant value outside the surface.

§ 2. THE SMOLUCHOWSKI EQUATION

Consider a spherical cell of radius R, volume V_0, in which the potential energy is a function of the distance from the centre. The mean number density of the liquid is accordingly $(N/V)=(1/V_0)$. The probability distribution of the molecule in the cell, to be denoted by $g(r, \theta, \phi)$ is in thermal equilibrium, given by

$$g=g_0=(1/Z).\exp(-U/\boldsymbol{k}T)$$

$$Z=\int \exp(-U/\boldsymbol{k}T)r^2 \sin\theta \, dr \, d\theta \, d\phi.$$

A non-equilibrium distribution which is independent of the time may be written

$$g = g_0(1+w),$$

where the function w is determined by the Smoluchowski equation (Chandrasekhar, 1943)

$$\operatorname{div}(\exp(-U/\boldsymbol{k}T) . \operatorname{grad} w) = 0, \qquad \ldots\ldots(1)$$

and is related to the flow vector according to

$$\mathbf{j} = (-\boldsymbol{k}T/\beta m) . \operatorname{grad} w. \qquad \ldots\ldots(1\,a)$$

In this equation m is the mass of the molecule and β is the "friction constant" of dimension $(\text{time})^{-1}$.

In the present theory equation (1) is assumed to hold in the interior of the cell even when the temperature or macroscopic velocity of the liquid is inhomogeneous. The effects of the departure from equilibrium are accounted for by conditions which the function w has to satisfy on the surface of the cell and according to which the density and the normal component of flow are made to be continuous on the surface. It should be said at once that these conditions can be satisfied only by functions with a singularity in the cell.

In the case of laminar flow in the direction of the y axis and of the rate of shear $2a$, the postulate of continuity of the normal component of flow requires that for $r=R$

$$(\boldsymbol{k}T/m) . g_0 . (\partial w/\partial r) = (ar/2NV_0) . \sin^2\theta . \sin 2\phi.$$

The factor $1/NV_0$ on the right-hand side accounts for the magnitude of the mean flow of one molecule. This condition is satisfied by putting

$$w = u(r) . \sin^2\theta . \sin 2\phi \qquad \ldots\ldots(2)$$

and demanding that

$$du/dr = [(armZ/2N\boldsymbol{k}TV_0) . \exp(U/\boldsymbol{k}T)], \qquad (r=R). \qquad \ldots\ldots(3\,a)$$

Continuity of density is possible only if

$$u = 0, \qquad (r=R). \qquad \ldots\ldots(3\,b)$$

If there is a gradient of temperature

$$T = T_0(1 + br . \cos\theta)$$

the density at the surface is determined by

$$g_0(1+w) = (1/NV_0) . \exp(-U(1-br . \cos\theta)/\boldsymbol{k}T_0)$$

or, writing T for T_0,

$$g_0 . w = br . \cos\theta . (U/\boldsymbol{k}T)(1/NV_0) . \exp(-U/\boldsymbol{k}T),$$

which is satisfied by putting

$$w = v(r) . \cos\theta \qquad \ldots\ldots(4)$$

and demanding that

$$v = brZU/NV_0\boldsymbol{k}T, \qquad (r=R). \qquad \ldots\ldots(5\,a)$$

When the flow of heat is steady there is no flow in the liquid, so that the normal component of flow must vanish at the surface of the cell:

$$\mathrm{d}v/\mathrm{d}r = 0, \qquad (r=R). \qquad \ldots\ldots(5\,b)$$

It will be seen that these surface conditions are decisive for the effect of temperature on the transport coefficients.

§ 2. SOLUTION OF THE EQUATION

The potential energy is, in the following, assumed to be

$$U = \tfrac{1}{2} G r^2,$$

so that the model requires altogether four constants for its complete specification, the mass m, the friction constant β, the force constant G and the cell radius R.

Introducing a dimensionless variable,

$$\xi = (G/2kT)^{\frac{1}{2}} . r,$$

the equilibrium distribution is

$$g_0 = (1/Z) . \exp(-\xi^2), \qquad Z = (2\pi kT/G)^{3/2}.$$

In order to find a solution of equation (1) which is appropriate to laminar flow, equation (2) is introduced into (1). From this a differential equation of u as function of ξ is obtained:

$$\xi^2 u'' + 2\xi(1-\xi^2)u' - 6u = 0, \qquad \ldots\ldots(6)$$

and solved by standard methods. The general solution is

$$u = A_1 u_1 + A_2 u_2,$$

where A_1 and A_2 are constants and

$$\left.\begin{aligned} u_1 &= \left[(1/\xi^3) . \exp \xi^2 . \int_0^{\xi} \exp(-t^2)\, dt\right] - (1/\xi^2) - (2/3), \\ u_2 &= -\left[(1/\xi^3) . \exp \xi^2 . \int_{\xi}^{\infty} \exp(-t^2)\, dt\right] - (1/\xi^2) - (2/3). \end{aligned}\right\} \qquad \ldots\ldots(7)$$

Similarly a solution of equation (1) is found which is appropriate to inhomogeneous temperature. Equation (4) is introduced into (1) so that a differential equation for v is obtained:

$$\xi^2 v'' + 2\xi(1-\xi^2)v' - 2v = 0. \qquad \ldots\ldots(8)$$

It has the general solution

$$v = B_1 v_1 + B_2 v_2,$$

where B_1 and B_2 are constants and

$$\left.\begin{aligned} v_1 &= \left[(2 + 1/\xi^2) . \int_0^{\xi} \exp t^2\, dt\right] - [(1/\xi) \exp \xi^2], \\ v_2 &= 2 + (1/\xi^2). \end{aligned}\right\} \qquad \ldots\ldots(9)$$

The functions u_1, v_1 increase exponentially for large values of ξ. The functions u_2, v_2 are finite at infinity and have a pole at $\xi = 0$. Since the density of force remains finite and the integral over w vanishes, there is no necessity for rejecting this calculation in spite of the singularities.

In the following calculations the expressions for the functions u and v are simplified in order to avoid lengthy formulae. The following first terms of expansions are used:

for $\xi \gg 1$

$$\begin{aligned} u_1 &= \tfrac{1}{2}\sqrt{\pi} . (1/\xi^3) \exp \xi^2 & u_1' &= \sqrt{\pi} . (1/\xi^2) \exp \xi^2 \\ u_2 &= -2/3 & u_2' &= 2/\xi^3 \\ v_1 &= (1/\xi^3) \exp \xi^2 & v_1' &= (2/\xi^2) \exp \xi^2 ; \end{aligned}$$

for $\xi \ll 1$

$$u_2 = -\tfrac{1}{2}\sqrt{\pi}\,.\,[(1/\xi^3)+(1/\xi)+(\xi/2)].$$

The constants of integration are determined by inserting the expressions for u and v into equations (3) and (5). In this calculation the expansions for large values of ξ are used. The result is

$$\left.\begin{aligned} A_1 &= 3\beta ma/4NG, \\ A_2 &= \frac{(9\sqrt{(2\pi)}\beta ma)\,.\,(\boldsymbol{k}T)^{3/2}\,.\,\exp(GR^2/2\boldsymbol{k}T)}{4NG^{5/2}R^3}; \end{aligned}\right\} \qquad \ldots\ldots(10)$$

$$\left.\begin{aligned} B_1 &= \frac{(3\sqrt{\pi b\boldsymbol{k}T})\,.\,\exp(-GR^2/2\boldsymbol{k}T)}{4NRG}, \\ B_2 &= 3\sqrt{\pi}(\boldsymbol{k}T)^{1/2}b/2^{3/2}NG^{1/2}. \end{aligned}\right\} \qquad \ldots\ldots(11)$$

The constants A_1, B_1 are clearly of smaller magnitude than the constants A_2, B_2. The constant A_2 contains an exponential factor $\exp(GR^2/2\boldsymbol{k}T)$ which appears in the coefficient of viscosity. The constant B_2 has no exponential factor.

§ 4. THE TRANSPORT COEFFICIENTS

The distribution functions u, v are used for calculating the flow of momentum and the flow of energy. In a dilute gas these quantities are determined by the amount of momentum and kinetic energy which isolated molecules carry along their path. In dense gases the transfer of momentum and energy during collisions is appreciable (Chapman and Cowling, 1939); the same mechanism of transfer prevails in liquids and solids, although there are no "collisions". Momentum and energy are continually exchanged between all those molecules for which the force of interaction is appreciable. Roughly, the flow of momentum is given by the negative product of force and distance, the flow of energy by the product of potential energy and relative velocity plus product of force, relative velocity and distance.

The flow of momentum and energy can be calculated according to formulae given by Kirkwood (1946). These formulae depend upon the distribution of two molecules in coordinates and moments, but can be rewritten in such a way that they apply to the cell model for which the distribution in coordinates only is given. This is obvious for the flow of momentum, which is equal to the stress produced by intermolecular forces. The flow of energy depends explicitly upon the momentum, but only in such a way as to be proportional to the flow of one molecule relative to another molecule. This flow can be derived from the distribution in relative coordinates, according to equation (1 *a*).

Denoting the orthogonal components of the position vector by s_i and of the flow vector by j_k $(i, k = 1, 2, 3)$, the components of the stress tensor are

$$P_{ik} = (N/2V_0)\,.\,\overline{[(\mathrm{d}U/\mathrm{d}r)(s_i\,.\,s_k)(1/r)]}$$

and the density of energy flow is

$$Q_i = (N/2V_0)\,.\,(\overline{Uj_i - [\tfrac{1}{2}(\mathrm{d}U/\mathrm{d}r)\textstyle\sum_k(s_i j_k + s_k j_i)s_k]}),$$

where the bar indicates the average over the distribution function.

The shearing stress is accordingly

$$P_{xy}=(N/2V_0)\iiint g_0\,.\,(\mathrm{d}U/\mathrm{d}r)ur^3\sin^5\vartheta\sin^2 2\phi\,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\phi.$$

The components of energy flow are

$$Q_r=(N/2V_0)\iiint [U-(r\,\mathrm{d}U/\mathrm{d}r)]j_r r^2\sin\theta\,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\phi,$$

$$Q_\theta=(N/2V_0)\iiint [U-\tfrac{1}{2}(r\,\mathrm{d}U/\mathrm{d}r)]j_\theta r^2\sin\theta\,\mathrm{d}r\,d\theta\,\mathrm{d}\phi,$$

$$Q_z=Q_r\cos\theta-Q_\theta\sin\theta.$$

The components of flow are

$$j_r=-g_0(\boldsymbol{k}T/\beta m)(\mathrm{d}v/\mathrm{d}r)\cos\theta,$$
$$j_\theta=g_0(\boldsymbol{k}T/\beta m)(v/r)\sin\theta,$$

so that

$$Q_z=(N\boldsymbol{k}T/2\beta mV_0\iiint (g_0[(r\,\mathrm{d}U/\mathrm{d}r)-U)(\mathrm{d}v/\mathrm{d}r)\cos^2\theta$$
$$+(\tfrac{1}{2}r\,(\mathrm{d}U/\mathrm{d}r)-U)(v/r)\sin^2\theta]r^2\,.\,\sin\theta\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\phi.$$

In calculating P_{xy} and Q_z the terms in A_1 and B_1 are neglected, the function u_2 is represented by its expansion for low values of ξ, and integration over the radius is extended to infinity. The result may be formulated in terms of the coefficient of viscosity η and the coefficient of thermal conductivity λ:

$$P_{xy}=-2a\eta=-2a(27/40\sqrt{(2\pi)})[m\beta(\boldsymbol{k}T)^{5/2}/R^6G^{5/2}]\exp(GR^2/2\boldsymbol{k}T), \quad\ldots\ldots(12)$$

$$Q_z=-bT\lambda=-bT(3/8\pi)(\boldsymbol{k}^2T/m\beta R^3). \quad\ldots\ldots(13)$$

§ 5. CONCLUSION

The coefficients of viscosity and thermal conductivity are thus obtained in terms of the friction constant. If this is supposed to depend only slightly upon the temperature, the two formulae give a reasonable representation of the type of temperature dependence of these two coefficients. Whereas the model and particularly the parabolic potential-energy curve cannot be expected to provide quantitative agreement with experiment, it is important to note that the reason for the different temperature effects is to some extent independent of the special assumptions. The only significant difference between the calculation of viscosity and thermal conductivity is in the boundary conditions. Their physical significance can be said to be the fact that molecules are forced to travel over regions of high potential energy in viscous flow, but not in a gradient of temperature. In the latter case they only carry out a circulation within the cell which distinguishes the case of a temperature gradient from that of equilibrium conditions.

This difference in mechanism between thermal conduction and viscous flow is independent of the present model and likely to be the explanation for the observed behaviour.

ACKNOWLEDGMENT

The writer is indebted to Professor E. N. da C. Andrade, F.R.S., for his interest in this work and for valuable discussions.

REFERENCES

ANDRADE, 1934. *Phil. Mag.*, **17**, 497, 698.
BORN and GREEN, 1947. *Nature, Lond.*, **159**, 251 ; 1946. *Proc. Roy. Soc.*, A, **188**, 10.
CHANDRASEKHAR, 1943. *Rev. Mod. Phys.*, **15**, 1.
CHAPMAN and COWLING, 1939. *The mathematical theory of inhomogeneous gases* (Oxford).
GRAETZ, 1885. *Ann. d. Phys., Lpz.*, **18**, 336.
GREEN, 1947. *Proc. Roy. Soc.*, A, **189**, 103.
KIRKWOOD, 1946. *J. Chem. Phys.*, **14**, 180.
KRAMERS, 1940. *Physica*, **7**, 284.

DISCUSSION

on paper by Dr. R. E. SIDAY, entitled "The Optical Properties of Axially Symmetric Magnetic Prisms" (*Proc. Phys. Soc.*, **59**, 905 (1947)).

Dr. O. KLEMPERER. Everybody who has worked practically with deflecting fields in velocity analysis or in spectrographs will have felt the great need which Dr. Siday has satisfied by attacking the problem of the magnetic prism from the theoretical side. The ballistic or electrodynamical approach is very complicated, and it implies great advantages to be able to treat the deflecting fields in terms of the optical methods. When I tried, many years ago, to apply the principles of design of the optical prism-spectrograph to β-ray spectroscopy, I was surprised to find the lines of β-spectra at all. However, these lines, or rather patches, were so broad that the lack of definition almost upset the high theoretical resolving power of the new instrument. Dr. Siday's method of cancelling the positive spherical aberration of the lenses by a negative aberration of the prism seems very promising. I should like to ask how far it will be possible to match the cylinder optics of the prism with the spherical optics (circular symmetry) of the lenses ? Does Dr. Siday expect that the lines of his spectra will be really straight or very much curved ?

Dr. H. H. HOPKINS. It will be interesting to see the results of Dr. Siday's investigation of rays not in the plane of symmetry, but, irrespective of what he finds for these rays, the prism that he has devised should still prove very useful. If the optical analogy holds, one could use a rectangular aperture with its longer dimension parallel to the plane of symmetry to permit only the corrected rays to form the image, and this without serious loss of resolving power in the direction perpendicular to the line-images.

AUTHOR's reply. Both Dr. Klemperer and Dr. Hopkins raise points of considerable importance, which will be considered in detail in Part II of this paper. In anticipation I would reply to Dr. Klemperer that by limiting the beam very straight lines can be obtained, and to Dr. Hopkins that I propose to use the prism in the way he suggests and I agree that over a limited aperture the resolution should not be seriously reduced.

DISCUSSION

on papers by C. GURNEY entitled "Delayed Fracture in Glass" (*Proc. Phys. Soc.*, **59**, 167 (1947)), and R. A. SACK entitled "Extension of Griffith's Theory of Rupture to Three Dimensions" (*Proc. Phys. Soc.*, **58**, 729 (1946)).

Prof. G. I. FINCH referred to the experimentally established existence of the cracks. He attributes their failure under load to adsorption of gases (H_2O, O_2, CO_2 in particular) towards the edge (end) of the crack exerting a wedging action. He drew attention to the adsorption of such gases on the fresh mica cleavage surface and the effect which such adsorption has on the re-packing of cleavage faces.

Mr. W. C. Hynd. Can the author say if cracks have been observed by the etching method adopted by Andrade on freshly blown samples of glass as distinct from mechanically polished surfaces ?

Mr. J. H. Awbery. The general idea of Griffith's argument is so appealing that one would be surprised if it did not agree with the closer analysis presented by atomic theory. Are they really independent, however ? The Griffith theory gives its criterion in terms of E and γ, but does not predict their values. When the observed values are inserted, are we not inserting the results which would follow from a successful atomic analysis of the phenomena ?

Mr. S. M. Cox. To subscribers to the Griffith flaw theory, Mr. Gurney's paper must supply a great deal of food for thought, but to a non-subscriber it seems rather to illustrate the expedients to which one is driven in order to account for the known facts on the basis of a theory which, in my view, begs the question.

I hope to show in a forthcoming paper that if one treats the problem from the beginning—that is to say, before the formation of the flaw instead of after it—and assumes that, under the conditions of test, the formation of such a flaw consitutes the trigger action for the rupture, then, at least for glass, all the salient facts are readily explained. The fact that rupture always progresses with a finite velocity I take to indicate that the theoretical strength (so called) is never actually reached.

The outstanding points are :

(1) The comparative weakness of the surface.

(2) The considerable spread of results under apparently identical testing conditions.

(3) The influence of time in producing eventual breakage without apparently weakening the glass before the breakage occurs.

(4) The apparent increased strength of fine fibres.

Taking into consideration the thermal energy of the ions together with the contribution due to the applied stress, the probability of the formation of a flaw of w ions may be shown to be proportional to

$$T\left\{N\left(e^{\frac{-B-S}{k\theta}}-2e^{\frac{-B+S}{k\theta}}\right)\right\}^{w},$$

where B is the potential barrier energy and S is the contribution to the total vibrational energy made by the applied tension. T is the duration of the test and N is the total number of ions.

If this expression does in fact define the condition for rupture, then it follows at once that :

(*a*) The surface is much weaker than the bulk except where the sample is exceedingly large, because the number of ions required to form a surface flaw is only about half that required for a non-surface flaw.

(*b*) The spread of the experimental results using one fiducial constant (viz. the mean strength) is given more closely than by a Gaussian distribution using two.

(*c*) There is excellent agreement with Preston's " Duration of Test " experiments.

(*d*) The strength of fine fibres should increase with diminishing radius as shown by Griffith's results but the agreement is only good down to about 0·05 mm. radius.

Where surface flaws, atmospheric attack, etc., are bound to play their part in practice, I believe the underlying explanation of the test results lies in the formation of flaws and not in their pre-existence.

Mr. R. W. Douglas. Experiment shows that the surface of glass has a very profound effect on its strength, particularly if it contains macroscopic cracks. Macroscopic cracks are easily formed on glass surfaces by subjecting the surface to some treatment, at a temperature a hundred degrees or so above room temperature, which removes alkali from the surface, e.g. attack by wet SO_2. This reduces the thermal expansion of the surface layers and on cooling the stresses set up by the differential contraction result in a system of very fine cracks. It is possible that a similar process in the ordinary manufacture of glass might produce such cracks on a microscopic scale. It is improbable therefore that effects such as

Mr. Cox has discussed can account for all the phenomena exhibited in the brittle rupture of glass. However, I agree with him in doubting that the Griffith crack mechanism is the only operative one.

X-ray diffraction experiments show that the arrangements of the atoms in a glass are exactly similar to those in a liquid if one could imagine all the translational motion in the liquid stopped. Now Temperley (*Proc. Phys. Soc.*, **58**, 436, 1946) has found that the tensile strength of water is much less than the strength predicted from theory. It is difficult to see how the Griffith theory can be applied to water and one wonders whether in glasses and in water there is not some other mechanism controlling rupture, and in view of the random atomic arrangements, whether it is possible that the estimated theoretical values of the strength are too high.

Mr. H. A. Elliott. Dr. Sack suggested that the normal Griffith criterion appeared to differ from that given by the condition that the stress at the apex of the crack should be equal to the theoretical strength of the material. I have shown (Elliott, 1947), however, that the theoretical strength condition leads to the relation $T \simeq P\sqrt{a/c}$, where P=theoretical strength, T=applied stress, c=semi-length of crack and a=lattice spacing. Also the surface energy, $S \simeq P^2 a/E$ where E is Young's modulus (this would hold on dimensiona grounds for a true law of atomic force as well as for the approximate one used in the quantitative calculation). Thus we see that $T \simeq \sqrt{ES/c}$ as Griffith finds (the numerical factor is relatively unimportant as it is of order unity) ; in the case of Griffith's experiments agreement is found numerically to approximately 10% error.

I agree with Gurney's statement that it seems unlikely that evaporation at the bottom of a crack is the cause of the slow fracture of glass under small loads. It should be observed that such a mechanism gives a law $t \simeq A\left(1 - e^{-\lambda\frac{T^2}{P^2}}\right)$, which does not agree even in form with the result of Preston (1946). It may, however, be possible to find a mechanism due to chemical attack (with water vapour) at the apex, the rate being mainly controlled by a diffusion process through the gel or alkali salts formed after attack. So far as I am aware this problem has not been solved. Dr. Orowan (1944) suggests a similar mechanism, quoting the fracture of mica. He also points out that the Griffith formula gives the correct strengths for both rapid and slow fracture if the vacuum value of the surface energy is used in the first case and the air-surface value in the second.

Author's reply. In reply to Mr. Hynd, Andrade (1937) found very few cracks on etching freshly blown glass, but they appeared on etching glass which had been exposed to air for some hours after blowing. This result is consistent with the origin of the cracks suggested in the paper, but no doubt there are other possible explanations.

In reply to Mr. Awbery, although the Griffith theory does not predict the values of E and γ, the values of these constants are not chosen so as to make the predictions of the Griffith theory agree most closely with experiment. For low stresses E is readily obtained by direct measurement, as is γ at high temperatures. The values inserted in the Griffith formula should be the values predicted by extrapolation of these direct measurements. The calculation of E and γ depends on quantum mechanics whereas the Griffith theory uses only classical ideas. In a more complete theory therefore it seems likely that the division of the subject into the prediction of E and γ and the utilization of these and other macroscopic quantities in a theory of fracture would still be convenient.

With regard to the remarks of Mr. Cox, I had thought that the activation energy necessary for the spontaneous formation of flaws was too high compared with the thermal energy at room temperature for the process he describes to be an appreciable factor in the strength of glass. If such an effect were predominant in causing delayed fracture it would seem that the theoretical strength of glass rods might be approached at very rapid rates of loading—a prediction which could perhaps be tested by experiment.

In reply to Mr. Douglas, a more recent paper by Temperley (1947) reduces the ratio of experimental to theoretical tensile strength of water to about 1 : 10. This ratio would probably be further reduced if Temperley's experiments could be repeated using more rapid rates of stressing. If an important delayed fracture effect were found for water a theory on the lines of that outlined by Mr. Cox might be applicable.

With regard to Mr. Elliot's remarks, I agree that the form of my equation (13) does not give a very good fit to Preston's results, and that the rate of diffusion of atmospheric constituents through the complex of glass and atmospheric constituents formed at the base of the crack may be an important factor in determining the time to fracture. The argument I put forward for the non-contamination of the material at the end of the crack during evaporation (p. 178 of the printed paper) would not apply to processes of chemical attack.

In conclusion I would like to emphasize that, although Griffith's theory may give results of the right order of magnitude, it ignores the time taken for the crack to spread. In solids there is very appreciable delay in the attainment of states of lower free energy, often so much so, that states having free energy above the minimum possible persist indefinitely. It is no more correct to assume that failure occurs immediately free energy considerations are favourable than it is to assume that glass will crystallize or evaporate in the same circumstances.

REFERENCES

Andrade, A. N. da C., 1937. *Proc. Roy. Soc.*, A, **159**, 346.
Elliott, H. A., 1947. *Proc. Phys. Soc.*, **59**, 208.
Orowan, 1944. *Nature, Lond.*, **154**, 341.
Preston, 1946. *J. Appl. Phys.* (March).
Temperley, H. N. V., 1947. *Proc. Phys. Soc.*, **59**, 199.

DISCUSSION

on paper by J. C. Evans entitled "The Determination of Thermal Lagging Times" (*Proc. Phys. Soc.*, 59, 242 (1947)).

Mr. D. K. Ashpole. Although Dr. Evans' paper deals primarily with the thermal lagging times of mercury barometers, reference is made to other possible applications. It therefore seems worth pointing out that care is necessary in applying the formulae to cases in which convection is important.

In the above paper it is assumed that the heat reaching any point in a column of liquid enclosed in a hollow cylinder is conducted either through the cylinder or through the liquid. This is only strictly true for an infinitely viscous liquid (i.e. a solid) due to the finite convection currents caused by the temperature gradient. Convection, however, will be negligibly small in comparison with conduction if the thermal conductivity of the liquid is high and the bore of the cylinder small. This accounts for the good agreement obtained between theory and practice for the mercury column defined in § 3 B.

For liquids other than mercury, the discrepancy between the theoretical and observed lagging times may become considerable. A case in point is a glass cylinder of 8·65 cm. bore, 0·35 cm. wall thickness, and 10 cm. long, containing water. A value of λ of 200 min. is obtained from Evans' equation (3), p. 246, namely,

$$\lambda_r = \frac{\rho_g S_g}{4k_g}(r_2^2 - r_1^2) + \frac{\rho_m S_m - \rho_g S_g}{2k_g} r_1^2 \log_e \frac{r_2}{r_1} + \frac{\rho_m S_m}{4k_m}(r_1^2 - r^2),$$

where ρ, S and k are the density, specific heat and thermal conductivity respectively, and the suffix g denotes the wall material, and the suffix m the liquid medium. The value $r=0$ was taken since the temperature was measured experimentally by means of a thermometer at the axis of the cylinder. The value of the lagging time obtained experimentally in a similar manner to that described by Evans was 9 min. In this instance, therefore, the convection is of greater moment than the conduction. Indeed, a better estimate of the lagging time is obtained by assuming perfect stirring within the cylinder (which gives a lower limit for λ) when a calculated value of 1 min. is obtained.

The applications of the equations derived by Evans should therefore be limited to cases where the effect of convection in the liquid can be shown or safely assumed to be negligible in comparison with conduction through it.

In the light of this conclusion, the value given on p. 248 for λ_0 with toluene in a thin-walled glass tube needs correction.

Author's reply. The theory given in the paper applies only to those cases in which convection effects are negligible, so that the results it gives will be erroneous when this condition is violated. Mr. Ashpole's note expresses the condition clearly.

OBITUARY NOTICES

FRIEDERICH PASCHEN

By a communication from his daughter, we have learned with deep regret that Professor F. Paschen died in Potsdam on 26 February 1947, at the age of 82. He was probably the greatest experimental spectroscopist of his time, and his record of discoveries is imposing both in its variety and in its importance. Fifty-five years have passed since Paschen's first classical publications on infra-red spectra appeared, and as the years progressed, his contributions to spectroscopy became more and more fundamental.

His earliest important publication on infra-red spectroscopy appeared in 1892. In the following year the extremely sensitive Paschen galvanometer was invented, and using it with a bolometer, Paschen was able to produce a deflection of 1 mm. for a temperature change of only 10^{-6} °c., a considerable achievement at the time. In 1894 he published the now classical determinations of infra-red refractive indices and dispersions for fluorite, rock-salt and quartz, and his values are still quoted in tables. This field was to interest him for several years to come, and although most of his energies were devoted to infra-red studies, yet it is of interest to note that as far back as 1895 he made measurements on the visible series spectrum of the newly discovered helium.

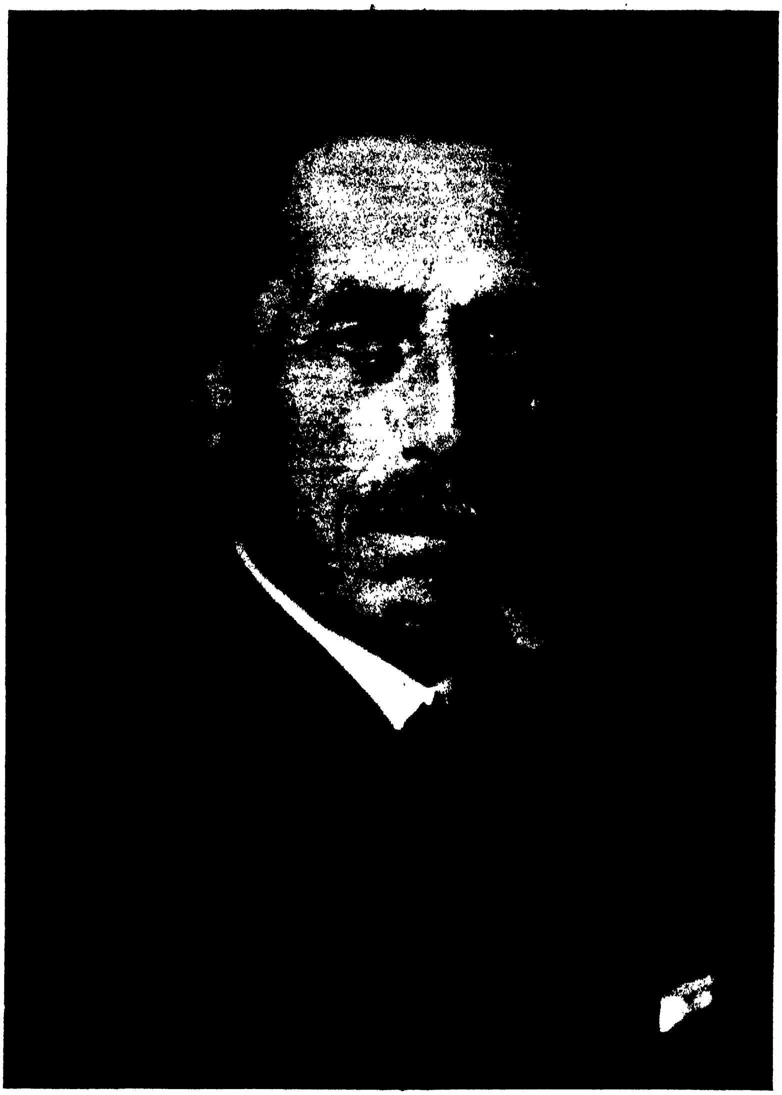
In 1899 Paschen founded the precision study of spectral black body radiation, and it is not widely enough realized that his pioneering work (which was followed by that of Lummer and Pringsheim) laid the experimental basis for the checking of Planck's quantum radiation formula. It was in this research that Paschen first introduced the use of concave mirrors for infra-red spectrographs, thus avoiding the considerable chromatic difficulties then inherent in the use of lenses. This technique is, of course, now absorbed into spectrography.

Paschen was always in the front line of progress, and in 1902 switched his attention to the investigation of the Zeeman effect, his studies of which were later to prove very fruitful indeed. His activities now ranged widely. In 1908 he found the infra-red Paschen series in hydrogen, so valuable later to the Bohr theory, and in 1912 the discovery of the Paschen-Back effect in magneto-optics led to the removal of many anomalies and clarified the whole study of the Zeeman effect. The year 1916 saw, in a sense, the beginning of the study of hyperfine structure, for Paschen then investigated the structure of the ionized helium line 4686; and his data proved to be critically important in the early history of the Bohr theory of the atom, affording also a precision value for **e/m**. It was at this time, too, that Paschen invented the hollow cathode discharge, later to prove so fruitful for hyperfine structure studies in the hands of his pupils.

In 1919 appeared, what is now classic in spectroscopy, Paschen's classification of 130 series in the spectrum of neon. This was the first complex spectrum to be analysed and revealed a masterly grasp of a difficult subject. Three years later, Paschen-Götz's comprehensive report on spectral series and lists of classifications appeared. Simultaneously, Fowler's report on " Series in Line Spectra " was published by our Society, and both books were standards of reference for many years. Fowler and Paschen, although widely differing in temperament, were comparable as masters in elucidating complex spectra and their work overlapped. Both discovered isoelectronic sequences of stripped atoms, Fowler with silicon and Paschen with aluminium, and both announced their discovery in 1923. Paschen's spectroscopic notation was perhaps slightly better than that used by Fowler, and more closely resembles modern practice.

The next ten years brought a spate of publications, on series, Zeeman effect, thermal excitation, hyperfine structure, nuclear spin etc. despite very heavy administrative duties; for during this period he was the President of the Physikalisch-Technische Reichsanstalt, and this was no sinecure.

From 1931 onwards Paschen was violently attacked in the Nazi press, and he confided that this was due to his persistent purchase of scientific equipment abroad, for he always demanded the best of its kind. He was dismissed in 1933 and replaced by Stark. After 1933 he was permitted to use a small room in the Reichsanstalt and contented himsef, with measuring old plates and analysing them. From then on publication dwindleld but this was certainly not due to the weakening of old age, for the portrait shown here,



FRIEDERICH PASCHEN.

PAUL LANGEVIN

taken after 70, shows his virility. Withal he remained a strong, if silent, opponent of the Nazis, and in 1936 wrote a bitter letter in which he declared that spectroscopy in Germany was now destroyed, and lived only in England, U.S.A. and India.

In 1943 he lost practically all his possessions in a fire-raid, but despite his age kept on working. In the extreme cold of January 1947, accentuated by lack of fuel, he contracted pneumonia and died on 26 February in Potsdam.

Paschen was a man with immense vitality, and a much-loved photograph of him, appearing in the *Annalen der Physik* on his sixtieth birthday, shows a man simply exploding with energy. He made it a habit to devote at least an hour daily to discussion with research students, despite an overload of administration. That he was a great experimenter is clear from his record, but what was equally important was his infectious enthusiasm. All of his pupils, and they are many, were infected by his adventurous attack at a problem, and even at 70 he was still as excited as a boy, either by a new photograph, a pretty technical achievement, or a fundamental result.

He had a great respect for English spectroscopists, in particular, for A. Fowler and was a profound admirer of the late Lord Rayleigh, about whose writings in the *Encyclopaedia Britannica* he frequently talked. He was a Foreign Member of the Royal Society and an Honorary Fellow of the Physical Society.

He was a just man and very helpful to his pupils, in fact, an ideal director of research, inspiring, suggesting, criticizing and guiding, yet never taking any credit for his contributions to the discussion despite their real importance. He was lavish in praise of good work, and frankly condemned bad work with an expressive contemptuous sweep of the hand. He has trained a whole generation of spectroscopists who have much to thank him for.

S. TOLANSKY.

PAUL LANGEVIN

PROFESSOR LANGEVIN died peacefully in the early hours of Thursday morning 19 December 1946 at his residence in the École de Physique et de Chimie Industrielle, of which he was Director. He had been ill for some days, and in view of his age—he was 74—his death was not unexpected. His loss will be deeply felt in France, not only on account of his great services to science but also on account of his devotion to patriotic and democratic ideals.

Paul Langevin was born at Montmartre, Paris, on 23 January 1872, his parents being of Norman peasant stock. At the age of 16 he entered the École de Physique et de Chimie Industrielle, and after winning the diploma of this institute he became a student at the École Normale Supérieure. When he ended his studies there, in 1897, the city of Paris awarded him a scholarship to pursue his work at the Cavendish Laboratory at Cambridge. He remained there a year, under Sir J. J. Thomson. On his return to Paris he became demonstrator first at the École Normale and then at the Sorbonne, and joined in research work with Pierre and Marie Curie. He took his doctorate of science in 1902, in which year he became assistant professor at the Collège de France; in 1909 he was elected to the chair of general and experimental physics. He succeeded Pierre Curie in the School of Physics and Chemistry, of which he became director in 1925. In 1934 he became a member of the Académie des Sciences.

His research in physics, the scope of which was immense, included work on x rays, the properties of ions, the kinetic theory of gases, molecular orientation, and magnetic theory. It was while working at Cambridge under J. J. Thomson that he discovered the secondary rays of x rays. He was the first in France to make known Einstein's formula of relativity. His work on the mobility of ions and his identification of the large ions in the atmosphere is classic. So, too, is his investigation of magnetism on molecular theory, which was extended by Weiss. To the general public he was chiefly known for his discovery of supersonic waves and their application to the detection of submarines, work in applied physics inspired by the needs of war; but his influence on French physicists was greater than his actual work would suggest. It was in this connection that he studied and utilized the piezo-electric properties of quartz.

Langevin was keenly interested in the Solway Institute of Physics, a body whose efforts in stimulating the progress of theoretical physics between the two wars was of outstanding influence. On the death of H. A. Lorentz of Holland he succeeded to the Presidency of this Institute.

Langevin had long been interested in politics and was known for his advanced ideas. In 1939 he was the moving spirit of *La Pensée*, a serious Marxist quarterly, so that when he returned to Paris in October 1940, whence he had withdrawn with other officials of the Ministry of National Defence, to which he was temporarily attached, he was arrested by the Germans, who sent him to the Fresnes concentration camp. He was the first prominent figure of the French learned world to be incarcerated by the Germans. Protests from intellectuals in many countries led to his release; he was allowed to take up his work at Troyes, under close supervision. Meanwhile his son-in-law, Jacques Salomon, was arrested and shot by the Germans, and his daughter, Mme Salomon, was deported to Germany. In May 1944 his friends in the Resistance planned his escape to Switzerland. He came back to Paris in September of the same year, and at once joined the Front National and the Communist Party. In December he became president of the *Ligue des Droits de l'Homme*, and in April 1945 was elected municipal counsellor for Paris. Soon afterwards, the Ministry of National Education appointed him chairman of the commission for the reform of educational methods, and in June of this year scientific advisor to the commissariat for atomic energy. Recently he was chosen to be deputy chief French delegate to the UNESCO conference.

JOHN HENRY STRONG

In the early hours of 13 April 1947 disaster overtook the aircraft carrying the British observers to Araxa, Brazil, for the total eclipse of the Sun on May 20. Among those who sustained fatal injuries was J. H. Strong, Demonstrator in Astrophysics and Spectroscopy at the Imperial College of Science and Technology and a Student Member of the Physical Society. By this tragic accident at Dakar was cut short what promised to be a most fruitful scientific career.

Strong was born in London on 17 August 1924 and was educated at the Latymer Upper School, Hammersmith. Awarded a State Bursary in 1942, he came to Imperial College and graduated with first class honours in physics in 1944. He was then appointed a Demonstrator in the Physics Department and began research in spectroscopy. His first investigation, a study of the properties of a type of arc source designed for general spectroscopy analysis, was proceeding very satisfactorily and he had planned the publication of some of his results after returning from Brazil.

In performing an experiment, Strong was unusually observant and would undoubtedly have become an experimentalist of outstanding ability. His manner was quiet and gave little outward sign of the active and methodical way in which he was building up his knowledge of his chosen field of research. Both as an instructor and as a research colleague he was greatly liked and respected. His loss is deeply regretted.

R. W. B. PEARSE.

REVIEWS OF BOOKS

A Survey of the Principles and Practice of Waveguides, by L. G. H. Huxley. Pp. xi+328. (Cambridge University Press, 1947). 21*s*.

During the war a great many fundamental advances were made in radio techniques as a result of the very rapid development of radiolocation (or radar as it now seems to be called) as a military weapon. These advances have been made known to the scientific public through the very successful Radiolocation Convention held in March last year under the auspices of the Institution of Electrical Engineers, and also through the published papers which have since appeared in the special issues (Part III A) of the Journal.* To many,

---

* Six special issues have now been published covering the lectures given at the Convention, and the supporting papers dealing with navigation and marine radar, aerials, waveguides, valves and cathode-ray tubes and receivers and transmitters.

however, these papers have been too specialized for easy study, and there has been a real need for a book, or books, providing an introductory survey of these developments. A series of books to fulfil this purpose has now been planned by the Cambridge University Press, and the present volume by Dr. Huxley dealing with the theory and practice of waveguides is the first of these to appear.

Dr. Huxley is Reader in Electromagnetism in the University of Birmingham, but, during the war, he was Head of the Radar School at the Telecommunications Research Establishment. He had thus a unique opportunity of becoming fully acquainted with all the recent developments, as well as learning at first hand the difficulties and pitfalls which present themselves in teaching the subject. We would therefore expect Dr. Huxley to write a clear and accurate account of the theory and practice of waveguides, in which the students' difficulties are appreciated and dealt with, and in this we are not disappointed. The treatment of the subject follows that of Dr. Huxley's courses at the Radar School. He deals with the properties and applications of waveguides, using simple physical explanations, and he relegates the formal mathematical treatment of field theory to the last chapter. The book appears to have lost nothing in content and vigour through this elementary and physical approach to the subject, and will certainly be appreciated by the less mathematical readers.

Chapter 1 gives an elementary treatment of the electromagnetic fields of T.E.M. waves. Chapter 2 discusses the general features of electromagnetic waves in metal waveguides and considers the various types of T.E. or H-waves and T.M. or E-waves and their relation to the T.E.M. wave. The chapter also includes a section dealing with the practical problem of launching these modes in circular and rectangular guides. Chapter 3 discusses the material of the earlier chapters in more mathematical language and evolves the formulae for the field components as functions of position in the guide. A discussion is also given of evanescent modes, piston attenuators, and attenuators of other types, as well as a consideration of attenuation due to the walls of the waveguides.

Chapters 4, 5 and 6 are decidedly the most interesting and probably the most valuable in the book. Chapter 4 discusses waveguide techniques, and deals with the special problems of the measurement of standing waves, the design of standing-wave indicators and their component parts (crystal detectors, crystal mixers etc.) reflectionless terminations, couplings, bends and corners, twists and tapers, and so on. All this material is well presented with full detail. § 4.12 dealing with $E_{0,1}$-H_0 transformers is a little unsatisfactory. The reviewer would have liked to have seen a much fuller account of these transformers.

Chapter 5 gives a general account of transmission line circuit theory using the concepts of characteristic impedance, line impedance, reflection coefficient etc., and then develops a parallel theory of waveguide transmission introducing the important concepts of waveguide impedance and normalized impedance. In all cases, however, where Dr. Huxley appeals to circuit theory, he is careful to point out the need for field concepts if a faithful description of the phenomenon is the aim. A physical interpretation of the normalized admittance or impedance of an obstacle or other discontinuity in a waveguide in terms of a scattering coefficient is given which should help in the understanding of these phenomena, and this is followed by a very full treatment of capacitive, inductive and resonant irises, etc. The treatment of T-junctions and their application to common aerial working T.R. systems, is excellent, but the section on waveguide switches is somewhat disappointing. The ring reflector switch and T.R. switch are quite rightly given special prominence, but little or nothing is said of other types, particularly variations of the ring switch, etc. The chapter also gives details of power measurements, the design of quarter-wave transformers and the design of corrugated waveguides. This latter is well presented, but very few details are given of the uses of corrugated waveguides and this section could well be amplified in the next edition. The chapter ends with a short section dealing with methods of feeding microwave aerials from waveguides. One slightly annoying feature of this and the previous chapter is the author's habit of occasionally referring by name to the originator of a particular idea or method. It would probably have been better to have omitted all reference to names of individuals (except, of course, when referring to published papers), or to have tried to mention all the chief contributors to the work. As it is, the author has omitted to mention the origin of many of the methods and techniques described, although the origin of many minor pieces of work is given in full.

Chapter 6 deals with cavity resonators and applications, and Chapter 7 attempts a formal treatment of field theory using single component Hertz vectors. This latter chapter contains much of interest and will well repay careful study.

The book is well printed and illustrated and appears to contain surprisingly few errors and mis-statements. It can be confidently recommended. D. TAYLOR.

Dissociation Energies and Spectra of Diatomic Molecules, by A. G. GAYDON, D.Sc. (Lond.), Warren Research Fellow of the Royal Society. Pp. xi + 239. (London: Chapman and Hall, Ltd., 1947.) 25*s*. net.

Dissociation energies of diatomic molecules, i.e. the energies required to dissociate such molecules from states without energy of translation, rotation or vibration, into states of two separate atoms at rest, are of the greatest theoretical interest, because they are the best measure of the chemical forces which bind two atoms together to form a molecule. They are also of great technical interest, e.g. in the investigation of combustion and of explosions.

The investigation of molecular spectra is a good example of the ever increasing co-operation of physicists and chemists. They can combine their efforts not only in the production of good spectra of pure materials, but also in the deciphering of the molecular spectra which are most complicated, and in their application to thermochemistry.

Dr Gaydon who has first edited an important book, *The Identification of Molecular Spectra*, with Dr. Pearse, and later an interesting book, *Spectroscopy and Combustion Theory*, publishes his new book from the standpoint of a thermochemist. He does not intend to introduce the reader to the theory of molecular spectra. This would be quite superfluous as we have an excellent introduction for physicists by G. Herzberg (*Molecular Spectra and Molecular Structure*) and for chemists by S. Glasstone (*Theoretical Chemistry*).

Gaydon gives only a very brief summary of the theoretical points which are relevant to the thermochemist (including such subjects as Potential Energy curves, Franck-Condon Principle and the correlations between molecular and atomic states), and concentrates mainly on his specific task to show first, by well investigated examples (O_2 and the halogen molecules), how the convergence limits of dissociation are found directly in molecular spectra and how the products of dissociation are determined. If the limits of convergence cannot be directly observed, this limit must be extrapolated. The method of extrapolation by Birge-Sponer is thoroughly discussed in Gaydon's book. The use which can be made of the phenomenon of predissociation for the determination of dissociation energies is treated in a separate chapter. Chapters are added on the thermal methods and the methods of controlled electron impacts which supplement the spectroscopic methods of determination of dissociation energies.

The most valuable part of the book consists in thorough *numerical* discussions of energies of dissociation of very important substances, especially of N_2, N_2^+, NO, CO, CO^+, CN, C_2N_2, and in a collection of the numerical data of about 250 diatomic molecules. Such critical data with all references from scientific periodicals and standard works can only be given by an expert and represent scientific work in its most condensed form.

The book which includes 39 figures and 4 plates of spectrograms, is written in a clear and stimulating style. It will be of great value for physicists as well as chemists (especially for thermochemists) and can be warmly recommended.

K. JELLINEK.

Fourier Transforms and Structure Factors, by DOROTHY WRINCH. Pp. ix + 96. (Asxred Monograph No. 2, 1946.) $4.00.

In a review of a scientific work it is often assumed that the reader is familiar with the subject matter, the reviewer restricting himself to the presentation of his views as to how well the author has succeeded in achieving his object, with hints, in some cases, on how it might have been done better. Occasionally, however, a book comes along on a subject with which the reader will almost certainly not be familiar, and in this case the reviewer's function becomes more that of an interpreter. Such a book is the monograph *Fourier Transforms and Structure Factors*, by Dorothy Wrinch, which arises from a long-term research on the structure of the native proteins, and which will certainly not be wholly understandable by every x-ray crystallographer, but which may well prove to be of great importance in structure analysis.

Briefly, Dr. Wrinch sets out to show how the structure factor of a particular atomic grouping may be combined with those of other atomic groupings to determine the structure factor of a crystal as a whole, so that a knowledge of the structure factors of frequently occurring atomic groupings may be used whenever the appropriate groupings arise, or are thought to be capable of arising, without recourse to the consideration of individual atomic positions. Difficulties on account of the variation of scattering power and atomic size are overcome by working throughout with the " transform " of a distribution (the structure factor per effective electron) rather than with the structure factor itself, and by plotting the values of these transforms in the reciprocal lattice on an arbitrary scale, the units of which can be varied appropriately to suit any particular case.

The transforms of a variety of frequently occurring groups—the vertices of a cube, of a tetrahedron and of an octahedron, and the "benzene" point set—are worked out, and it is shown how the transform concept can be extended to include Patterson distributions. From these relatively simple transforms Dr. Wrinch goes on to consider transforms of many points arranged in patterns on surfaces or in depth, and finally considers the transforms of various distributions in continuous space, rather than periodic distributions such as occur in crystals. A chapter is added on the application of the transform concept to the study of small crystals, and there is an appendix on Fourier series and Fourier's integral theorem.

In the preface, Dr. Wrinch makes the point that a systematic study of the " language of structure factors " is a necessary preliminary to the interpretation of the intensity maps of crystals made up of megamolecules of unknown structures. The present monograph will render such a study much easier and should certainly do something to encourage the use of transforms in ordinary crystal analysis, a use which has so far been very restricted.

J. THEWLIS.

Sound: A Physical Textbook, by E. G. RICHARDSON, D.Sc. 4th edition. Pp. 344. (Edward Arnold and Co., 1947.) 18*s*.

Since the appearance of the first edition of Dr. Richardson's book in 1927 much water has flowed under London Bridge, or perhaps one might say more appropriately " much sound has degenerated into heat ". Much of this sound served a useful purpose but we know full well that much of it did not!

In the preface to the first (1927) edition Dr .Richardson referred to the renewed attention to this branch of physics " stimulated no doubt by the European War (1914–18) and by the development of broadcasting ". If further stimulus were required this has no doubt been provided by the second war. Dr. Richardson's well known book has served a very useful purpose between the two wars in encouraging the study of that very interesting subject of " Sound ". His book has now reached the fourth edition after a period of 20 years since its first appearance. In the latest edition the whole subject matter has been re-arranged and brought up to date. It is fitting to quote from the preface of this edition: " The position in sound is now very different from what it was when this book was first written : there are a number of books dealing with special aspects of applied acoustics and at the outbreak of the second world war there were three periodicals and one society dealing solely with the subject, whilst in industry the penetration of applied acoustics has been far-reaching ". It may be of interest to add also that the Physical Society has recently formed an Acoustics Group which has already a very considerable membership both of " academic " and ' industrial ' scientists and engineers.

Dr. Richardson covers the general field of sound and, the reviewer is pleased to note, is now taking a little more interest in under-water sound. The chapters on " Vortex Formation and Jet Tones " and on " Columns of Air " are particularly noteworthy, as no other textbook of sound known to the reviewer deals so thoroughly with these important branches of the subject. They are all the more valuable as Dr. Richardson himself has played a leading part in the work which he describes. Referring again to under-water sound, there is a little confusion on p. 324 in the description of the construction and mode of operation of two hydrophones, one of the 'light-body', 'displacement type' (illustrated) and the other of the metal diaphragm 'pressure type'.

The book is very readable and covers a wide field in a relatively small compass. The appearance of the fourth edition is sufficient testimony to its continued popularity and usefulness. This excellent textbook needs no further recommendation. A. B. WOOD.

German-English Science Dictionary, by DE VRIES. Second edition, revised and enlarged; second impression. Pp. xiv+558. (London: McGraw-Hill Publishing Co., 1947.) 22*s*. 6*d*.

This is a dictionary intended for students in the physical, biological and related sciences. Originally containing 48,000 entries, this new edition contains additional words and idioms and incorporates a great number of suggestions made by users of the first.

The gender and plural form of each noun are given explicitly, except for nouns belonging to the large class of regular feminine plurals. One most welcome feature in a dictionary of this kind, which in the main has users who are not language specialists, is the printing of the parts of strong verbs as separate entries. It is thus not necessary, for example, to recognize *schlägt* or *schlug* as parts of the verb *schlagen* before one can find their meanings. The infinitive of the verb is given in brackets after each part so printed. This should prove of considerable assistance to students (and others) for whom languages are not a very strong point.

Compound nouns are, in general, not entered as such, the reader being expected to look up the different parts of such words separately. It must be said, however, that this rule is very often not adhered to, even in cases where the compounding has introduced no special meaning. Thus one finds *Messergebnis*, *Messfehler*, etc., entered separately—and such examples occur on almost every page. However, to most users this will be looked upon rather as an advantage.

A weakness, and one which tends to mar this otherwise excellent dictionary, is the failure to distinguish compound verbs in their separable and inseparable forms. This is serious, since the same verb used separably may differ completely in meaning from the inseparable form. Thus *über/legen*=to lay over; but used inseparably *überlegen*=to consider. This verb is, in fact, entered only once—and these two widely differing meanings are apparently given as synonyms ! It is to be hoped that this will be corrected in future editions.

This last criticism should not be allowed to detract from the fact that in other respects this dictionary offers real assistance to anyone who has occasion to read scientific German—and that includes a large enough body of workers to say that many people will find this work helpful and some invaluable. H. H. HOPKINS.

Guide to the Literature of Mathematics and Physics, by NATHAN GREER PARKE. III. Pp. xv+205. (New York and London: McGraw-Hill Book Co. Inc., 1947.) $5.00.

This book falls into two parts. About two-fifths is occupied with a general guide in which the respective uses of handbooks, text-books, encyclopedias and dictionaries are set out, methods of study discussed, methods of literature search described, indexing and cataloguing shortly treated, the main abstracting journals mentioned, and so on. This material is quite interesting and helpful, but differs entirely from that in the second part of the book. In the first part, just described, the material is for study and assimilation, whereas the second part is simply, and valuably, for reference.

The second part contains, in classified sections, the titles and bibliographical particulars of some 2300 books likely to be of use at one time or another to any physicist. The three main fields, of mathematics, physics and engineering, are each split up into a few broad areas (9 for physics) and each of these into perhaps a dozen sub-areas, of which there are 150 altogether. These 150 sub-areas are then given in alphabetical order (so that sub-areas from different broad ones come together) and under each alphabetical entry is given a short commentary and a list of text-books or sources which might be consulted on it.

These books are not confined to those from any one country; French and German, American and British, all appear. It would not be appropriate to discuss in detail the selection actually made—it is a personal one, and the reviewer's, whilst it might differ, would be equally personal. It may be of interest, however, to prospective purchasers, to know what sort of thing he is likely to find, and we may fairly note, therefore, that in some cases early sources are included, but that this is not always so. Thus, all the five books on algebraic equations are published in América and dated 1938 or later, so that Cayley, Sylvester and Salmon do not appear, and the treatise by Burnside and Panton is not listed. On the other hand, under " Elliptic functions ", we find the *fundamenta nova* of Jacobi (1829) but not Lagrange's *Traité*. Here, again, we find a standard English treatise, that of Cayley (1895) but not the very useful little book by Dixon.

It may, then, be seen that here is a useful reference book which may be very helpful when one is seeking the literature of a subject new to one, and may be a pleasant browsing companion at other times. It is as well produced as the McGraw books generally are.

J. H. A.

The Principles of Quantum Mechanics, by P. A. M. DIRAC. 3rd edition. Royal 8vo. Pp. 311. (Oxford: The University Press, 1947.) 25*s*. net.

It is not often that a book written within five years of the first introduction of an entirely new theory, so wide and so fundamental as quantum theory and wave mechanics, remains the standard textbook for more than fifteen years. Yet this is the case with Dirac's *Principles of Quantum Mechanics*. It was as long ago as 1930, less than five years after the introduction of Heisenberg's matrix mechanics, and Schrodinger's wave mechanics, that the first edition of this book appeared, and the late Sir Arthur Eddington wrote of it that "for the first time wave mechanics is presented in a really coherent form with something like a philosophy of the new methods to support it ". The fact that so few fundamental changes have been found necessary in later editions (of which this is the third) is proof enough that Eddington was right. Though the book is not everywhere easy to read, no one who is genuinely interested in the general formulation of quantum theory can afford to neglect it.

Substantially this third edition closely resembles the second edition. The theory of quantum electrodynamics has advanced a little (but only a little) since 1935, so that there is now found room for the Wentzel field, and the λ-limiting process. As Dirac says, to go further would be to trespass " on speculative ground ". A more powerful method of dealing with systems of like particles is used, based on Fock's treatment of the theory of radiation. But the most noticeable change, at least to the eye, is the introduction of the *bra* and *ket* notation. What was previously called ψ, the representative of a state, is now called a *ket*, and written $|\,\rangle$; the dual vector, hitherto called ϕ, is now called a *bra*, and written $\langle\,|$. The product of the two vectors $\langle B|$ and $|A\rangle$, which is always a pure number, and which appears very frequently, is no longer $\phi\psi$, but $\langle B|A\rangle$. The use of *bra* and *ket* is obvious when we decide to call this latter expression a " bracket ", or " complete bracket ", compared with which the others are " incomplete brackets ". The change of notation has the advantage of greatly simplifying some of the formalism.

As usual the Oxford University Press has turned out a first-rate piece of printing.

C. A. C.

Applied Bessel Functions, by F. E. RELTON. Pp. vi+191, 10 figures. (Blackie and Son, 1947.) 17*s*. 6*d*.

This book, written by an applied mathematician, is very unconventional in the order of presentation it adopts. The author is at pains to justify this approach in his preface—and draws on the support of E. B. Wilsòn in this connection. He had in fact little need of this, for his text provides more than adequate justification.

The material of the book is presented in a uniformly pleasant (almost chatty) style, which would perhaps be felt obtrusive in a work of bigger proportions. Here, however, in a book intended for the new physics or engineering graduate, it seems quite in place. What is more it enables the author to point a few morals to the tyro in applied mathematics which a more formal style might well have inhibited. As a consequence these interpolations make the book both more interesting and more informative.

The first chapter is devoted to a preliminary study of the error function, and of the gamma and beta functions, while Chapter 2 reviews those parts of the theory of linear differential equations which are necessary to later parts of the book. In Chapter 3 a class of functions is defined by means of recurrence formulae. It is shown that these functions are solutions of Bessel's equations, and are accordingly Bessel functions. The series solution of Bessel's equation is not given until Chapter 4; and Chapter 5 then deals with applications involving the Bessel functions of the first kind, to which attention is at first confined.

Chapters 6 and 7 see the introduction of the Bessel functions of the second kind, the modified Bessel functions, and further applications calling for their use. The applications to hydrodynamics and elasticity, to which Chapter 8 is devoted, will probably be found the most difficult for the beginner. The various integrals and expansions involving Bessel functions are not treated until Chapter 9—in itself sufficient to evidence the unconventional presentation of the subject. The final chapter deals very briefly with allied functions and asymptotic expansions.

For anyone anxious to know something about Bessel functions there are books as good as this, but few could claim to be better. H. H. HOPKINS.

Photographic Recording of Cathode-ray Tube Traces, by R. J. HERCOCK. Pp. 60. (London: Ilford Ltd., 1947.) 5*s*.

The book forms " No 1 of a series of Ilford technical monographs ", and " It is the object of this booklet to provide those familiar with the use of cathode-ray tubes with an insight into the photographic technique necessary to obtain useful records of traces ". The book does not, however, deal solely with the photographic aspects. It also describes tubes and methods of application. The author contrives to cover a wide field in the short space of sixty pages by adopting a pleasantly terse style and by approaching the reader at a fairly high technical level. Thus, certain photographic terms, such as " density " are used but not explained, as is indeed unnecessary. The book thus assumes some familiarity both with physical and photographic terms.

Photography as a scientific tool is treated as a Cinderella in some laboratories. For this reason, the book will form a welcome addition in the libraries of many institutions and will fill a gap among existing text-books.

The scope of the book does not extend beyond published material. Insufficient stress is laid on the necessity for full development with high recording speeds and thus short exposures; there is also no mention of recent work on latent image intensification, which is often useful in this field. There is one peculiar misprint: a density of 0·1 is quoted as corresponding to a change of 12½ % in transmission, the correct figure being 26 %. The book suffers but slightly from the naturally exclusive preoccupation with Ilford materials, and the scientific world will look forward to further volumes in this series.

W. F. B.

Light, Vision and Seeing, by MATTHEW LUCKIESH. Pp. xiv + 323. New York: D. Van Nostrand Company, Inc., 1944.) 25*s*. net.

In an earlier work—*The Science of Seeing*—Dr. Luckiesh and Mr. Moss recapitulated the valuable series of investigations on visual capacity, lighting and human welfare which had been carried out over some thirty years by the authors and their colleagues at the Research Laboratories of the General Electric Company, Cleveland. The present book, addressed to readers interested but not necessarily knowledgeable in the subject, gives a simplified and detailed exposition of the more elementary topics of the earlier work, states the principal conclusions and includes additional matter on the historical development of lighting methods and their impact on everyday life. While the English reader may find the style a little florid in places, there is no doubt that Dr. Luckiesh has produced an easily understood and very informative account of the relation between light and sight. There are a few slips of the pen, for example the statement on p. 287 that the molecules of air are comparable in size with the wave-lengths of light. W. S. S.

INDEX

PAGE

Absorption and selective photo-effect in adsorbed layers . . . 30

Acceleration of charged particles to very high energies . . . 666

Acceleration of electrons, Multiple-gap linear . . . 622

Adiabatic temperature changes and magnetization of cobalt . . . 329

Aircraft camera lenses; discussion (**58**, 493, 1946) . . . 155

Allen, W. D. and Symonds, J. L.: Experiments in multiple-gap linear acceleration of electrons . . . 622

Ammonia: Collision broadening of the inversion spectrum at centimetre wavelengths: I . . . 418

Andrews, J. P.: Thermoelectric power of cadmium oxide; and discussion . . . 990

Aplanatic lenses for unit magnification . . . 844

Appleton, Sir Edward, and Beynon, W. J. G.: The application of ionospheric data to radio communication problems: Part II; and discussion . . . 58, 534

Appleton, Sir Edward, and Naismith, R.: The radio detection of meteor trails and allied phenomena . . . 461

Arc, copper, The effect at the cathode . . . 273

Aspheric profiles, On the determination of . . . 704

Auditorium, Mean free path of sound in . . . 535

Avery, D. G. and Witty, R.: Diffusion pumps: A critical discussion of existing theories . . . 1016

β-spectroscopes, prism, The design of . . . 905

Band spectra of CS and CSe . . . 107

Band-spectroscopic data, The calculation of potential-energy curves from . . . 998

Band-systems, ultra-violet absorption, of PbO, PbS, PbSe and PbTe . . . 449

Bands of Na$_2$, Ultra-violet . . . 610

Barrow, R. F., *see* Vago, E. E.

Bate, A. E. and Pillow, M. E.: Mean free path of sound in an auditorium . . . 535

Bates, L. F. and Edmondson, A. S.: The adiabatic temperature changes accompanying the magnetization of cobalt in low and moderate fields . . . 329

Bates, W. J.: A wavefront-shearing interferometer . . . 940

Beynon, W. J. G.: Oblique radio transmission in the ionosphere, and the Lorentz polarization term; and discussion . . . 97, 534

Beynon, W. J. G.: Observations of the maximum frequency of radio communication over distances of 1000 km. and 2500 km.; and discussion . . . 521, 534

Beynon, W. J. G., *see* Appleton, Sir Edward.

Bleaney, B., Loubser, J. H. N. and Penrose, R. P.: Cavity resonators for measurements with centimetre electromagnetic waves . . . 185

Bleaney, B. and Penrose, R. P.: Collision broadening of the inversion spectrum of ammonia at centimetre wavelengths: I . . . 418

Bragg x-ray spectrometer, Optical model demonstrating . . . 111

Brossel, J.: Multiple-beam localized fringes . . . 224, 234

Bromine, Interpretation of the visible spectrum of . . . 1008

Brown, R. C.: The fundamental concepts concerning surface tension and capillarity; and discussion, and corrigenda . . . 429, 711

Brunt, Professor D., President of the Physical Society, 1945–1947; Portrait . . . *frontispiece*

Brunt, D.: Some physical aspects of the heat balance of the human body. (Presidential Address) . . . 713

Bryan, G. B. (Obituary notice) . . . 508

Burch, C. R.: Reflecting microscopes . . . 41, 47

Butler, C. C., *see* Rymer, T. B.

PAGE
Cadmium oxide, Thermoelectric power of 990
Camera lenses, Aircraft; discussion (58, 493, 1946) 155
Capillarity and surface tension, Fundamental concepts concerning; and corrigenda 429, 711
Cathode of the copper arc, A note on the effect at 273
Cavity resonators for measurements with centimetre electromagnetic waves . . . 185
Centimetre electromagnetic waves, Cavity resonators for measurements with . 185
Centimetre wavelengths, Collision broadening of the inversion spectrum of ammonia at: I 418
Centimetric electromagnetic waves over ground, Reflection and diffraction effects with wire-netting screens 847
Charged particles to very high energies, Acceleration of 666
CO, Photo-chemical decomposition of; addendum to discussion 502
Coaxial electron lenses 828
Cobalt, The magnetization and adiabatic temperature changes of 329
Collision broadening of the inversion spectrum of ammonia at centimetre wavelengths: I 418
Colorimeter with six matching stimuli 554
Colorimetry in the glass industry 592
Coloured light signals near the limit of visibility, Recognition of 560
Coloured point sources against a white background, Measurement of the chromatic and achromatic thresholds of 574
Colours, Distribution coefficients for the calculation of, on the C.I.E. trichromatic system for total radiators 814
Corrigenda 33, 711, 901
Coutts, W. B. (Obituary notice) 508
Cowley, J. M. and Rees, A. L. G.: Refraction effects in electron diffraction . . 287
Craggs, J. D. and Hopwood, W.: Electron/ion recombination in hydrogen spark discharges 771
Craggs, J. D. and Hopwood, W.: Ion concentrations in spark channels in hydrogen 755
Craig, H.: The production of a uniform magnetic field over a specific volume by means of twin conducting circular coils 804
Crystal structure of gold leaf, Determination by electron diffraction . . . 541
CS and CSe, The spectra of 107
Cubic lattice, Dislocations in a simple 256
Cuer, P., *see* Lattes, C. M. G.

Daly, E. F. and Sutherland, G. B. B. M.: An infra-red spectroscope with cathode-ray presentation; and discussion 77, 901
Damping capacity, strain hardening and fatigue. 275
Damping, radiation, theory of 917
Diffraction effects with wire-netting screens, and reflection of centimetric electromagnetic waves over ground 847
Diffusion, The hole theory of 694
Diffusion pumps: a critical discussion of existing theories 1016
Dislocations in a simple cubic lattice 256
Domb, C., The theory of an oscillator coupled to a long feeder, with applications to experimental results for the magnetron 958
Donaldson, R.: A colorimeter with six matching stimuli 554
Dungey, J. W. and Hull, Catherine R.: Coaxial electron lenses 828
Dymond, E. G.: The Kew radio sonde 645

Edmondson, A. S., *see* Bates, L. F.
Edmondson, A. S.: The matching of high frequency transmission lines using a frequency-variation method 982
Eisenschitz, R.: The effect of temperature on the thermal conductivity and viscosity of liquids 1030
Electro-acoustic transducers, Sensitivity and impedance of 19
Electromagnetic centimetre waves, Cavity resonators for measurements with . . 185
Electron diffraction, Refraction effects in 287

PAGE
Electron/ion recombination in hydrogen spark discharges 771
Electron lenses, Coaxial 828
Electron optics and space charge in strip-cathode emission systems 302
Electrons, Experiments in multiple-gap linear acceleration of 622
Elliott, H. A.: An analysis of the conditions for rupture due to Griffith cracks 208
Emissivity of hot metals in the infra-red 118, 131
Evans, J. C.: The determination of thermal lagging times; and discussion 242, 1039

Fatigue, damping capacity, strain hardening and 275
Faust, R. C. and Tolansky, S.: A transparent-replica technique for interferometry 951
Flicker noise in valves and impurity semi-conductors; and discussion 366, 403
Fluctuations in streams of thermal radiation 34
Fluctuations, spontaneous, of electricity in thermionic valves; and discussion 375, 388,403
Fowler, P. H., *see* Lattes, C. M. G.
Fracture in glass, Delayed; and discussion 169, 1036
Frank, F. C.: The mass of the neutrino 408
Freezing-in of nuclear equilibrium 139
Fröhlich, H. and Sack, R. A.: Light absorption and selective photo-effect in adsorbed layers 30
Fundamental physics, What experiments are needed in (lecture and discussion) 412
Fürth, R. and MacDonald, D. K. C.: Spontaneous fluctuations of electricity in thermionic valves; and discussion 375, 388, 403

Generator, High-voltage, at Imperial College 699
Gerö, L., *see* Schmid, R. F.
Glass, Delayed fracture in; and discussion 169, 1036
Glass industry, Colorimetry in the 592
Glückauf, E.: Investigations on absorption hygrometers at low temperatures 344
Gogate, D. V. and Pathak, P. D.: The Landau velocity in liquid helium : II 457
Gold leaf, Determination of the crystal structure by electron diffraction 541
Gooden, J. S., *see* Oliphant, M. L.
Gooden, J. S., Jenson, H. H. and Symonds, J. L.: Theory of the proton synchroton; and corrigenda 677, 901
Griffith's theory of rupture, Extension of, to three dimensions; discussion (**58**, 729, 1946) 1036
Griffith cracks, An analysis of the conditions for rupture due to 208
Grimmett, L. G., *see* Mann, W. B.
Gurney C.: Delayed fracture in glass; and discussion 169, 1036
Gurney, C.: Thermodynamic relations for two phases containing two components in equilibrium under generalized stress 629

Hamilton, J.: The theory of radiation damping 917
Hanstock, R. F.: Damping capacity, strain hardening and fatigue 275
Harding, H. G. W. and Sisson, R. B.: Distribution coefficients for the calculation of colours on the C.I.E. trichromatic system for total radiators at 1500–250–3500° K., and 2360° K. ($C = 14\,350$) 814
Heat balance of the human body, Some physical aspects of the 713
Heat transfer, Some investigations in the field of 726
Heavy elements in stars, The formation of 972
Helium II, liquid, Landau velocity in 457
Hey, J. S., Parsons, S. J. and Jackson, F.: Reflection of centimetric electromagnetic waves over ground, and diffraction effects with wire-netting screens 847
Hey, J. S. and Stewart, G. S.: Radar observations of meteors 858
Hide, G. S., *see* Oliphant, M. L.
High-frequency transmission lines, matching by a frequency-variation method 982
Hill, N. E. G.: The measurement of the chromatic and achromatic thresholds of coloured point sources against a white background 574
Hill, N. E. G.: The recognition of coloured light signals near the limit of visibility 560

PAGE
Hole theory of diffusion . . . 694
Holmes, J. G.: Colorimetry in the glass industry . . . 592
Hopwood, W., *see* Craggs, J. D.
Howell, H. G.: The spectra of CS and CSe . . . 107
Hoyle, F.: The formation of heavy elements in stars . . . 972
Hull, Catherine R., *see* Dungey, J. W.
Human body, Some physical aspects of the heat balance of . . . 713
Hydrogen, Ion concentrations in spark channels in . . . 755
Hydrogen spark discharges, Electron/ion recombination in . . . 771
Hydrostatic tension, The behaviour of water under, III . . . 199
Hygrometers, absorption, at low temperatures . . . 344

Impedance and sensitivity of electro-acoustic transducers . . . 19
Imperial College high-voltage generator . . . 699
Inclined plane pole-faces, Magnetic focusing between . . . 791
Infra-red spectroscope with cathode-ray presentation; and discussion . . . 77, 901
Infra-red, The emissivity of hot metals in the . . . 118, 131
Interferometer, A wavefront-shearing . . . 940
Interferometry, A transparent-replica technique for interferometry . . . 951
Intermittency effect, A note on the . . . 161
Ion concentrations in spark channels in hydrogen . . . 755
Ionosphere, Oblique radio transmission in, and the Lorentz polarization term; and discussion . . . 97, 534
Ionospheric data, The application to radio communication problems: II; and discussion . . . 58, 534
Ionospheric region, Equivalent path and absorption in an . . . 87

Jackson, F., *see* Hey, J. S.
Jaeger, J. C.: Equivalent path and absorption in an ionospheric region . . . 87
Jakob, Max: Some investigations in the field of heat transfer . . . 726
Jeans, Sir James (Obituary notice) . . . 503
Jensen, H. H., *see* Gooden, J. S.
Jets, liquid, The break-up of . . . 1

Kew radio sonde . . . 645
Klemperer, O.: Electron optics and space charge in strip-cathode emission systems 302

Laby, T. H. (Obituary notice) . . . 506
Lagging times, The determination of thermal . . . 242
Landau velocity in liquid helium II . . . 457
Langevin, P. (Obituary notice) . . . 1041
Latham, R.: Nuclear magnetic moments . . . 979
Lattes, C. M. G., Fowler, P. H. and Cuer, P.: A study of the nuclear transmutations of light elements by the photographic method . . . 883
Lenses, Aircraft camera; discussion . . . 155
Lenses, aplanatic, for unit magnification . . . 844
Lewis, W. B.: Fluctuations in streams of thermal radiation . . . 34
Lines of force through neutral points in a magnetic field . . . 14
Liquid helium II, Landau velocity in . . . 457
Liquid jets, The break-up of . . . 1
Liquids, The effect of temperature in the thermal conductivity and viscosity of . 1030
Lord, Mary P., Rees, A. L. G. and Wise, M. E.: The short-period time variation of the luminescence of a zinc sulphide phosphor under ultra-violet excitation; and corrigenda . . . 473, 711
Loubser, J. H. N., *see* Bleaney, B.
Luminescence of a zinc sulphide phosphor under ultra-violet excitation, Short-period time variation of; and corrigenda . . . 473, 711

MacDonald, D. K. C., *see* Fürth, R.
Macfarlane, G. G.: A theory of flicker noise in valves and impurity semi-conductors; and discussion . . . 366, 403

PAGE
Magnetic field, Lines of force through neutral points in 14
Magnetic field, Uniform production over a specific volume by means of twin conducting circular coils 804
Magnetic focusing between inclined plane pole-faces 791
Magnetic moments, nuclear 979
Magnetic prisms, The optical properties of axially symmetric 905
Magnetization of cobalt in low and moderate fields, The adiabatic temperature changes accompanying 329
Magnetron, Application of theory of an oscillator and long feeder to experimental results for 958
Mann, W. B. and Grimmett, L. G.: The Imperial College high-voltage generator. 699
Maximum frequency of radio communication over distances of 1000 km. and 2500 km.; and discussion. 521, 534
May, J.: corrigendum 33
Merrington, A. C. and Richardson, E. G.: The break-up of liquid jets . . 1
Meteor trails and allied phenomena, The radio detection of 461
Meteors, Radar observations of 858
Micrometer, time, of high accuracy 585
Microscopes, Reflecting 41, 47
Milbourn, M.: The effect at the cathode of a copper arc 273
Multiple-beam localized fringes 224, 234
Multiple-gap linear acceleration of electrons, Experiments in 622

Na_2, Ultra-violet bands of 610
Nabarro, F. R. N.: Dislocations in a simple cubic lattice 256
Naismith, R., *see* Appleton, Sir Edward.
Neutrino, The mass of 408
Neumann, E. A.: A time micrometer of high accuracy 585
Nickel, The variation of the reflectivity with temperature 781
Nuclear equilibrium, The freezing-in of 139
Nuclear magnetic moments 979
Nuclear transmutations of light elements, A study by the photographic method . 883

Obituary notices 503, 1040
Oliphant, M. L.: Rutherford and the modern world. (Third Rutherford lecture). 144
Oliphant, M. L., Gooden, J. S. and Hide, G. S.: The acceleration of charged particles to very high energies 666
Optical properties of axially symmetric magnetic prisms 905
Oscillator coupled to a long feeder, with application to results for the magnetron . 958
Owen, D.: The lines of force through neutral points in a magnetic field; and corrigendum 14, 901

Parsons, S. J., *see* Hey, J. S.
Paschen, F. (Obituary notice) 1040
Paterson, S.: The heating or cooling of a solid sphere in a well-stirred fluid . 50
Pathak, P. D., *see* Gogate, D. V.
PbO, PbS, PbSe and PbTe, Ultra-violet absorption band-systems of 449
Peierls, R. E.: What experiments are needed in fundamental physics? (lecture and discussion) 412
Penrose, R. P., *see* Bleaney, B.
Permittivity, effective, of two-phase systems 1011
Phosphor, zinc sulphide, Short-period time variation of the luminescence of; and corrigenda 473, 711
Photo-chemical decomposition of CO; addendum to discussion 502
Photo-effect and light absorption in adsorbed layers 30
Pillow, M. E., *see* Bate, A. E.
Preddy, W. S., *see* Wolf, E.
Price D. J.: The emissivity of hot metals in the infra-red 118, 131
Prisms, magnetic, axially symmetric, Optical properties of 905

PAGE
Profiles, aspheric, On the determination of 704
Proton synchroton, Theory of the 677
Pumps, diffusion: a critical discussion of existing theories 1016

Radar observations of meteors. 858
Radiation damping, The theory of 917
Radiators, total, Distribution coefficients for the calculation of colours on the C.I.E. trichromatic system for 814
Radio communication, the maximum frequency of, over distances of 1000 km. and 2500 km.; and discussion 521, 534
Radio communication problems: The application of ionospheric data to: II; and discussion 58, 534
Radio detection of meteor trails and allied phenomena 461
Radio sonde, The Kew 645
Radio transmission, oblique, in the ionosphere, and the Lorentz polarization term; and discussion 97, 534
Rees, A. L. G., *see* Cowley, J. M.
Rees, A. L. G., *see* Lord, Mary P.
Rees, A. L. G.: The calculation of potential-energy curves from band-spectroscopic data 998
Rees, A. L. G.: Note on the interpretation of the visible absorption spectrum of bromine 1008
Reflecting microscopes 41, 47
Reflection of centimetric electromagnetic waves over ground, and diffraction effects with wire-netting screens 847
Reflectivity of nickel, The variation of, with temperature 781
Refraction effects in electron diffraction 287
Replica, transparent, technique for interferometry 951
Resonators, cavity, for measurements with centimetre electromagnetic waves . 185
Reviews of books 157, 326, 509, 711, 902, 1042
Richardson, E. G., *see* Merrington, A. C.
Richardson, H. O. W.: Magnetic focusing between inclined plane pole-faces . 791
Rupture, Extension of Griffith's theory of, to three dimensions; discussion (**58**, 729, 1946) 1036
Rupture due to Griffith cracks, An analysis of the conditions for 208
Rushman, D. F. and Strivens, M. A.: The effective permittivity of two-phase systems 1011
Rutherford and the modern world (Third Rutherford lecture) 144
Rymer, T. B. and Butler, C. C.: Determination of the crystal structure of gold leaf by electron diffraction 541

Sack, R. A.: Extension of Griffith's theory of rupture to three dimensions; discussion (**58**, 729, 1946) 1036
Sack, R. A., *see* Fröhlich, H.
Schmid, R. F. and Gerö, L.: Addendum to discussion 502
Selwyn, E. W. H. and Tearle, J. L.: The performance of aircraft camera lenses; discussion (**58**, 493, 1946) 155
Semi-aplanat reflecting microscopes 47
Semi-conductors, impurity, A theory of flicker noise in valves and; and discussion 366, 403
Sensitivity and impedance of electro-acoustic transducers 19
Siday, R. E.: The optical properties of axially symmetric magnetic prisms: I; and discussion 905
Sinha, S. P.: Ultra-violet bands of Na_2 610
Sisson, R. B., *see* Harding, H. G. W.
Smith, T.: A series for the stationary value of a function 323
Smith, T.: Note on aplanatic lenses for unit magnification 844
Sound in an auditorium, Mean free path of 535
Space and charge in strip-cathode emission systems, Electron optics and . . . 302
Spark channels in hydrogen, Ion concentrations in 755

PAGE

Spark discharges, hydrogen, Electron/ion recombination in 771
Spectra of CS and CSe 107
Spectra, ultra-violet absorption, of PbO, PbS, PbSe and PbTe 449
Spectroscope, infra-red, with cathode-ray presentation; and discussion. . 77, 901
Spectrum of bromine, visible absorption, Interpretation of 1008
Spectrum of Na_2 610
Sphere, solid, The heating or cooling in a well-stirred fluid. 50
Stars, The formation of heavy elements in 972
Stationary value of a function, A series for the 323
Stewart, G. S., *see* Hey, J. S.
Strain hardening, Damping capacity and fatigue 275
Stress, generalized, Thermodynamic relations for two phases containing two components in equilibrium under 629
Strivens, M. A., *see* Rushman, D. F.
Strong, J. H. (Obituary notice) 1042
Surface tension and capillarity, Fundamental concepts concerning; and discussion 429, 711
Sutherland, G. B. B. M., *see* Daly, E. F.
Symonds, J. L., *see* Allen, W. D.
Symonds, J. L., *see* Gooden, J. S.
Synchrotron, proton, Theory of the 677

Tearle, J. L., *see* Selwyn, E. W. H.
Temperley, H. N. V.: The behaviour of water under hydrostatic tension: III . 199
Thermal conductivity and viscosity of liquids, effect of temperature on. . . 1030
Thermal lagging times, The determination of 242
Thermal radiation, Fluctuations in streams of 34
Thermodynamic relations for two phases containing two components in equilibrium under generalized stress 629
Thermoelectric power of cadmium oxide 990
Thermionic valves, Spontaneous fluctuations of electricity in; and discussion 375, 388, 403
Time micrometer of high accuracy 585
Tolansky, S., *see* Faust, R. C.
Transducers, electro-acoustic, Sensitivity and impedance of 19
Transmutations, nuclear, of light elements, A study by the photographic method . 883
Transparent-replica technique for interferometry 951

Ubbelohde, A. R.: The freezing-in of nuclear equilibrium 139
Uranium chain reaction, A mechanical model illustrating the 113

Vago, E. E. and Barrow, R. F.: Ultra-violet absorption band-systems of PbO, PbS, PbSe and PbTe 449
Valves and impurity semi-conductors, A theory of flicker noise in; and discussion 366, 403
Valves, thermionic, Spontaneous fluctuations of electricity in; and discussion 375, 388,403
Vigoureux, P.: Sensitivity and impedance of electro-acoustic transducers . . 19
Viscosity and thermal conductivity of liquids, Effect of temperature on. . . 1030

Ward, F. A. B.: A mechanical model illustrating the uranium chain reaction . 113
Ward, F. A. B.: A simple optical model demonstrating the principle of the Bragg x-ray spectrometer 111
Water under hydrostatic tension, The behaviour of: III 199
Wavefront-shearing interferometer 940
Weil, R.: A note on the intermittency effect 161
Weil, R.: The variation of the reflectivity of nickel with temperature . . . 781
Wise, M. E., *see* Lord, Mary P.
Witty, R., *see* Avery, D. G.
Wolf, E. and Preddy, W. S.: On the determination of aspheric profiles . . 704
Wyllie, G.: The hole theory of diffusion 694

X-ray spectrometer, Bragg, Optical model demonstrating 111

Zinc sulphide, Short period time variation of the luminescence of; and corrigenda 473, 711

INDEX TO REVIEWS OF BOOKS

PAGE
Aharoni, J.: *Antennae: An Introduction to their Theory* 328

Cady, W. G.: *Piezoelectricity* 513
Carslaw, H. S. and Jaeger, J. C.: *Conduction of Heat in Solids* 519
Cohen, B. S.: *A Handbook of Telecommunication* 158
Cosslett, V. E.: *Introduction to Electron-Optics* 327

De Vries: *German-English Science Dictionary* 1046
Dirac, P. A. M.: *The Principles of Quantum Mechanics* 1047

Frenkel, J.: *The Kinetic Theory of Liquids* 516

Gaydon, A. G.: *Dissociation Energies and Spectra of Diatomic Molecules* . . 1044
Gill, F. C.: *The Vector Operator j* 520
Graves, L. M.: *The Theory of Functions of Real Variables* 517

Hercock, R. J.: *Photographic Recording of Cathode-ray Tube Traces* . . . 1048
Huxley, L. G. H.: *A Survey of the Principles and Practice of Waveguides* . . 1042

Jaeger, J. C.: *see* Carslaw, H. S.
Jeffreys, B. S.: *see* Jeffreys, H.
Jeffreys, H. and Jeffreys, B. S.: *Methods of Mathematical Physics* 512

Luckiesh, Matthew: *Light, Vision and Seeing* 1048

McLachlan, N. W.: *Theory and Application of Mathieu Functions* 711
Mathematical Tables Project: *Tables of Fractional Powers* 518
Mathematical Tables Project: *Tables of Spherical Bessel Functions* . . . 520

Parke, Nathan Greer: *Guide to the Literature of Mathematics and Physics* . . 1046
Pidduck, F. B.: *see* Sas, R. K.
Pipes, L. A.: *Applied Mathematics for Engineers and Physicists* 515

Relton, F. E.: *Applied Bessel Functions* 1047
Richardson, E. G.: *Physical Science in Art and Industry* 517
Richardson, E. G.: *Sound: A Physical Textbook*. 1045
Russell, B.: *Physics and Experience* 509

Sas, R. K. and Pidduck, F. B.: *The Metre-Kilogram-Second System of Electrical Units* 903
Sommer, A.: *Photoelectric Cells* 157

Tolansky, S.: *Introduction to Atomic Physics* 520

Watson, W. H.: *The Physical Principles of Wave Guide Transmission and Antenna Systems* 902
Weatherburn, C. E.: *A First Course in Mathematical Statistics* 160
Willmer, E. N.: *Retinal Structure and Colour Vision* 518
Wright, W. D.: *Researches on Normal and Defective Colour Vision* 326
Wrinch, D.: *Fourier Transforms and Structure Factors* 1044

Young, V. J.: *Understanding Microwaves* 159

I.A.R.I. 75

INDIAN AGRICULTURAL RESEARCH INSTITUTE LIBRARY, NEW DELHI

Date of Issue	Date of Issue	Date of Issue
10.10.62		
30-7-64		

GIPNLK—H-40 I.A.R.I.—29-4- 5—15,000